Faculty for Electrical Engineering and Information Technology
Institute for Media Technology

Dissertation

A thesis submitted for the degree of
Doctor of Engineering

On the plausibility of simplified acoustic room representations for listener translation in dynamic binaural auralizations

Submitted by: **Dipl.-Ing. M.Sc. Annika Neidhardt**

Supervisor: Prof. Dr.-Ing. Dr. rer. nat. h.c. mult. Karlheinz Brandenburg
Reviewer: Prof. Dr.-Ing. Christoph Pörschmann
Reviewer: Prof. Dr. Stefan Weinzierl

Date of submission: 06th December 2022
Date of public defence: 25th May 2023

DOI: 10.22032/dbt.57596
URN: urn:nbn:de:gbv:ilm1-2023000155

Abstract

This thesis investigates the effect of simplified acoustic room representations in position dynamic binaural audio for listener translation.

Dynamic binaural synthesis is an audio reproduction method to create spatial auditory illusions over headphones for virtual, augmented, and mixed reality (AR/VR/MR). It has become a common demand to explore such illusionary contents in six degrees of freedom (6DOF). Realizing dynamic binaural sound field imitations with high physical accuracy is usually linked to high computational effort. However, previous psychoacoustic research indicates that humans have limited sensitivity to the details of the sound field, especially in reverberation. This fact bears the potential to simplify the physics in the position dynamic auralization of rooms. For example, concepts based on the perceptual mixing time or the audibility threshold of early reflections have been proposed, but lack a thorough psychoacoustic evaluation and validation for listener translation. Hence, this thesis investigates the effect of simplified acoustic room representation in position dynamic binaural audio for listener translation. The effects are measured in terms of the plausibility of the resulting spatial auditory illusion.

To pursue related investigations, first, a setup for position dynamic binaural room auralization was implemented and evaluated. Essential system parameters like the required position grid resolution for the audio reproduction were examined.

Furthermore, there was a lack of generally established test methods for the perceptual evaluation of spatial auditory illusions considering interactive listener translation. Consequently, this thesis explores different approaches to measuring plausibility and assessing potential degradations of plausibility caused by simplified rendering. Based on this foundation, this work examines physical impairments and simplifications in the progress of the sound field in position dynamic binaural auralizations of room acoustics.

For the main experiments, sets of binaural room impulse responses (BRIRs) were measured along a line for listener translation in a relatively dry listening laboratory and a considerably more reverberant seminar room of similar size. These sets include scenarios of walking towards a virtual sound source, past it, away from it, or behind it. Moreover, the investigations consider two extreme cases of source orientation to take into account the effects of a variation in the sound source directivity. The BRIR sets are systematically impaired and simplified to evaluate the perceptual effects. Especially the concept of the perceptual mixing time and manipulated spatiotemporal patterns of early reflections served as test cases in the psychoacoustic studies.

The results reveal a high potential for simplification but also underline the relevance of accurately imitating strong early reflections. The findings also confirm the concept of the perceptual mixing time for the considered cases of position dynamic binaural reproduction. The observations highlight that common test scenarios for auralizations, interpolation, and extrapolation are not sufficiently critical to draw general conclusions about the suitability of the tested rendering approaches. This thesis proposes approaches to solve this.

Zusammenfassung

Diese Dissertation untersucht die Effekte von vereinfachten akustischen Raumabbildungen in positionsdynamischem binauralen Audio für die Hörertranslation.

Die dynamische binaurale Synthese ist eine Audiowiedergabemethode zur Erzeugung räumlicher auditiver Illusionen über Kopfhörer für virtuelle, erweiterte und gemischte Realität (VR/AR/MR). Es ist zu einer gängigen Anforderung geworden, solche illusionären Inhalte in sechs Freiheitsgraden (6DOF) zu erkunden. Dynamische binaurale Schallfeldimitationen mit hoher physikalischer Genauigkeit zu realisieren, ist meist mit hohem Rechenaufwand verbunden. Frühere psychoakustische Studien weisen jedoch darauf hin, dass Menschen eine begrenzte Empfindlichkeit gegenüber den Details des Schallfelds haben, insbesondere im Nachhall. Diese Tatsache birgt das Potential, die Physik bei der positionsdynamischen Auralisation von Räumen zu vereinfachen. Beispielsweise wurden Konzepte vorgeschlagen, die auf der perzeptiven Mixing Time oder der Hörbarkeitsschwelle von frühen Reflexionen basieren, denen jedoch eine gründliche psychoakustische Bewertung und Validierung fehlt. Daher untersucht diese Arbeit die Effekte einer vereinfachten akustischen Raumrepräsentation in positionsdynamischem binauralem Audio für Hörertranslation. Die Wirkung wird anhand der Plausibilität der entstehenden räumlichen auditiven Illusion gemessen.

Um entsprechende Untersuchungen durchzuführen, wurde zunächst ein Aufbau zur positionsdynamischen binauralen Raumauralisation implementiert und evaluiert. Untersucht wurden wesentliche Systemparameter wie die erforderliche räumliche Auflösung eines Positionsrasters für die dynamische Anpassung.

Darüber hinaus fehlten allgemein etablierte Testmethoden zur wahrnehmungsbezogenen Bewertung von räumlichen auditiven Illusionen unter Berücksichtigung interaktiver Hörertranslation. Daher untersucht diese Arbeit verschiedene Ansätze zur Messung der Plausibilität und zur Bewertung möglicher Verschlechterungen der Plausibilität, die durch vereinfachtes Rendering verursacht werden. Auf dieser Grundlage untersucht diese Arbeit physikalische Beeinträchtigungen und Vereinfachungen im Verlauf des Schallfeldes in positionsdynamischen binauralen Auralisationen der Raumakustik.

Für die Hauptexperimente wurden Sätze von binauralen Raumimpulsantworten (BRIRs) entlang einer Linie für die Hörertranslation in einem relativ trockenen Hörlabor und einem wesentlich halligeren Seminarraum ähnlicher Größe gemessen. Diese Sätze beinhalten Szenarien von Hörerbewegungen auf eine virtuelle Schallquelle zu, daran vorbei, davon weg oder dahinter. Darüber hinaus betrachten die Untersuchungen zwei Extremfälle der Quellorientierung, um die Auswirkungen einer Variation der Schallquellenrichtcharakteristik zu berücksichtigen. Die BRIR-Sätze werden systematisch bearbeitet und vereinfacht, um die Wahrnehmungseffekte zu bewerten. Insbesondere das Konzept der perzeptiven Mixing Time und manipulierte räumlich-zeitliche Struktur früher Reflexionen dienten als Testfälle in den psychoakustischen Studien.

Die Ergebnisse zeigen ein hohes Potential für Vereinfachungen, unterstreichen aber auch die Relevanz der genauen Nachahmung starker früher Reflexionen. Die Ergebnisse bestätigen auch das Konzept der wahrnehmungsbezogenen Mixing Time für die betrachteten Fälle der positionsdynamischen binauralen Wiedergabe. Die Beobachtungen verdeutlichen, dass gängige Testszenarien für Auralisierungen, Interpolation und Extrapolation nicht kritisch genug sind, um allgemeine Schlussfolgerungen über die Eignung der getesteten Rendering-Ansätze zu ziehen. Die Arbeit zeigt Lösungsansätze auf.

Acknowledgements

The past few years have been an intense time for me. I had the chance to explore the science world on my own behalf, which was very exciting and tempting on the one hand, but also accompanied by frustration and self-doubts, on the other hand. Luckily, I had people that helped me through difficult times, and I am super-grateful for having them particularly in the times of need.

First, I want to thank Karlheinz Brandenburg for giving me the opportunity to pursue a PhD, for the freedom to let several of my ideas become reality, which let me grow as a scientist. I also thank him for his persistent faith in me throughout the good and the bad days.

I thank Christoph Pörschmann, who has always had an open ear (and a mouth that kept emitting sound with a certain directivity ;)) for all the challenges I was facing on my way to submission day.

I thank Stefan Weinzierl for the feedback and ideas on how to improve my data analysis.

I want to thank my group, especially the "old rabbits" Florian, Stephan, and Ulrike, for sharing the boat with me - Together, we have weathered several storms, set up lots of gorgeous research proposals, let an ICSA happen, let a book chapter happen, let the group grow, survived many students - good ones and uninterested ones, survived many students' projects and presentations - super-good ones and super-bad ones, set up a research colloquium for praising the 'fruitful' in the discussion, more or less survived university bureaucracy, and I think, we actually managed to enhance the state of the art in dynamic binaural audio research ... well, if that isn't something :)

Thank you Moni, for your incredible patience and support. Thanks Matthias, Torsten, Thomas, Bernd, Christiane for all your help! Thanks Hörbert for never giving up on me.

I absolutely want to thank my students Maria, Afrooz, Samaneh, Alexandra W., Alexandra D., Tahereh, Avinash, Nawres, Martina, Tatiana, Nuzhat, Sadia, Josephin, Julia, Chenyao, Shuang (Sorry Gentlemen, but Ladies first!), Niklas, Anson, Alby, Christian, Robin, Berni, Stefan, Manan, Marcel, Nils, Throni, Kai, Alvaro, Tarek, Kenneth, Oliver, Sebastian, Michael, David, Lokesh, Philip, Boris, Aravindan, Jean Claude, Mhd Ammar, Oliver, Georg, Tobias H., Tobias B., Bibek - Thank you for all your hard work, your ideas, your creativity, your persistence, for all that I have learned from you and for your incredibly inspiring personalities. It was such a pleasure to meet you and such an honor to work with you!!!

I ultimately thank Rakesh, Steve and Tatiana for successfully pushing video-streaming knowledge into my head before my Rigorosum, for lovely evenings with lovely food and lovely conversations. That thanks, of course, also goes to Felix and Chenyao. Rakesh, no worries - the mission is on - soon, we will swim in money :) Steve, thank you for the nicest door bell sign ever :) Oh, yes and thank you for the lovely SK storybook :)

Thank you so much, Chrissi, for all that shared office time, for lots of fun, lots of gossip, all the inspiring TDWs, for lots of fruitful discussions and actual research, for the chess matches, the balcony conversations, honest opinions, and for a view from the far! You will tell me, when it's time for SA? I am definitely in! :) And of course I hope that this is not the end.

Thank you, Frank, for the zoom-room and all that was shared within. It has made each of these days brighter for me, especially the dark ones!

I definitely need to thank Sirko for tasting with me through the W-menu while discussing the up- and downsides of the world with me. How would I have survived these years without this valuable input??? :)

Christin, thanks for sharing that bottle of wine with me!!
Anne, thanks for making me run! And for being the same, no matter where and when we meet!
Thank you, Micha, for everything and still liking me! It means a lot to me!

Thanks to my family for actually being my family! I love every single one you!!!
Thank you, M., for taking me to the mountains, to the waves, to the music, to the lakes, to the radio, the interview, the restaurant, the barbecue, to the snow, to the flow, to the sun, to the fun, to the Ziege, to the Zirbe, to the dome, and home. :) I carry it all in my heart!

List of Publications - Selection related to the thesis

1. **A. Neidhardt**, C. Schneiderwind, and F. Klein, "Perceptual matching of room acoustics for auditory augmented reality in small rooms - a review", Trends in Hearing, Vol. 22, 2022. https://doi.org/10.1177/23312165221092919

2. **A. Neidhardt** and A. M. Zerlik, "The availability of a real hidden reference affects the plausibility of position-dynamic auditory AR," Frontiers in Virtual Reality, 2021. https://doi.org/10.3389/frvir.2021.678875

3. K. Brandenburg, F. Klein, **A. Neidhardt**, U. Sloma, and S. Werner, "Creating auditory illusions with binaural technology", in The Technology of Binaural Understanding. Springer Int. Publishing, 2020, pp. 623-663. https://doi.org/10.1007/978-3-030-00386-9_21

4. R. Gupta, J. He, R. Ranjan, W.-S. Gan, F. Klein, C. Schneiderwind, **A. Neidhardt**, K. Brandenburg, V. Välimäki, "Augmented/Mixed Reality Audio for Hearables: Sensing, control and rendering", IEEE Signal Processing Magazine, Vol. 39, No. 3, May 2022. https://doi.org/10.1109/MSP.2021.3110108

5. **A. Neidhardt** and B. Reif, "Minimum BRIR grid resolution for interactive position changes in dynamic binaural synthesis," in 148th AES Convention, paper 10371, Vienna, Austria, 2020. - **Best Student Paper Award.**

6. **A. Neidhardt**, "Effect of incorrect early reflections on plausibility of position-dynamic binaural AR audio and its limits," unpublished, 2023.

7. **A. Neidhardt**, A. Ignatious-Tommy, and A. D. Pereppadan, "Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets," in 144th AES Convention, paper 9987, Milan, Italy, May 2018.

8. **A. Neidhardt** and S. Kamandi, "Plausibility of an approaching motion towards a virtual sound source II: In a reverberant seminar room," in 152nd AES Convention, paper 10608, The Hague, The Netherlands / Online, 2022.

9. C. Schneiderwind, **A. Neidhardt**, and D. Meyer, "Comparing the effect of different open headphone models on the perception of a real sound source," in 150th AES Convention, paper 10489, Online, May 2021.

10. S. Werner, F. Klein, **A. Neidhardt**, U. Sloma, C. Schneiderwind, and K. Brandenburg, "Creation of Auditory Augmented Reality using a position-dynamic binaural synthesis system - Technical components, psychoacoustic needs, and perceptual evaluation," Applied Sciences 11(3), Jan., p. 1150, 2021. https://doi.org/10.3390/app11031150.

11. **A. Neidhardt**, F. Klein, N. Knoop, and T. Köllmer, "Flexible python tool for dynamic binaural synthesis applications," in 142nd AES Convention, e-Brief 346, Berlin, Germany, 2017.

12. **A. Neidhardt** and N. Knoop, "Binaural walk-through scenarios with actual self-walking using an HTC Vive," in 43rd Annual Conference on Acoustics, Kiel, Germany, 2017.

13. **A. Neidhardt**, "Data set and physical analysis: BRIRs and SRIRs for walking toward, past and behind virtual loudspeakers in two rooms," in 154th AES Convention, Espoo, Finland, 2023.

14. **A. Neidhardt**, "Data set of measured room impulse responses: BRIRs, RIRs, SRIRs for position-dynamic binaural auralization in two rooms," Data set (1.0), Zenodo, Tech. Rep., 2023. zenodo.org/record/7838178

15. L. Remaggi, K. Hansung, **A. Neidhardt**, A. Hilton, and P. B. Jackson, "Perceived quality and spatial impression of room reverberation in VR reproduction from measured images and acoustics," in 23rd Int. Congress on Acoustics, Aachen, Germany, 2019.

16. **A. Neidhardt**, "Perception of the reverberation captured in a real room, depending on position and direction," in 22nd Int. Congress on Acoustics (ICA), Buenos Aires, Argentina, 2016.

17. **A. Neidhardt**, "Data set: BRIRs for position-dynamic binaural synthesis measured in two rooms," in 5th Int. Conference on Spatial Audio, Ilmenau, Germany, 2019.

18. **A. Neidhardt**, A. M. Zerlik, and S. Kamandi, "BRIR data set for interactive listener translation in two rooms," Data set (1.0), Zenodo, Tech. Rep., 2020. zenodo.org/record/3457782

19. K. Brandenburg, E. Cano Ceron, F. Klein, T. Köllmer, H. Lukashevich, **A. Neidhardt**, J. Nowak, U. Sloma, and S. Werner, "Personalized Auditory Reality", in 44th Annual Meeting on Acoustics (DAGA), Garching (Munich), Germany, 2018.

20. **A. Neidhardt**, K. P. Jurgeit, A. Nasrollahnejad, and J. Nowak, "Investigating continuous adaptation of binaural reproduction to changing listening position" Annual Meeting on Acoustics (DAGA), Garching (Munich), Germany, 2018.

21. **A. Neidhardt** and C. Schneiderwind, "The influence of the DRR on audiovisual coherence of a real loudspeaker playing virtually over headphones," in 47th Annual Meeting on Acoustics (DAGA), Vienna, Austria, 2021.

Contents

1. Introduction

Listening with two ears, humans can distinguish different sound sources, localize them in direction and distance, determine the acoustic attributes of the surrounding room and understand complex auditory scenes. For spatial hearing, the brain analyzes the signals arriving at the left and the right ear and compares them. For example, interaural time differences (ITD) and interaural level differences (ILD), as well as spectral cues, are known as the primary acoustic cues in the estimation of the source direction [1]. Auditory distance perception is more complex, especially in rooms. Besides the absolute and relative sound level, the relation of the sound energy arriving directly from the sound source and the reflections arriving from all directions (direct-to-reverberant-energy-ratio - DRR) is an important cue to estimate the distance of a sound source [2, 1]. Other factors, like spectral information, familiarity with the sound source, and visual cues, influence auditory distance perception. Understanding the relevant acoustic cues for spatial hearing and how the human auditory sense processes them is fundamental to binaural synthesis, which is subject to research in this thesis.

1.1. Binaural Synthesis

Binaural synthesis is an audio reproduction method that attempts to mimic the acoustic cues humans use for spatial hearing like ILD, ITD, spectral information, or DRR at both ears. The goal is to create spatial auditory illusions of single sound objects or whole acoustic environments. Binaural reproduction was developed to be realized over headphones [3]. Furthermore, loudspeaker systems that are optimized for delivering the signals for the left and the right ear with minimum crosstalk can be used [4, pp. 283-325]. The basic idea is to produce an approximation of the sound pressure at both eardrums, which is sufficiently similar to that created by a corresponding real version of the sound object or environment. Creating an exact duplicate of that sound pressure is hardly possible due to technical limitations. Evidently, it is also not necessary because the human auditory system cannot make use of all the detailed information. In order to realize binaural reproduction inducing auditory illusions of high quality efficiently, it is crucial to study how the human auditory system analyzes the signals at both ears and which acoustical information is necessary. One of the main challenges in this regard are cognitive effects. How the human brain interprets the ear signals does not only depend on the physical properties of these signals but also on the individual experiences and expertise, expectations, the actual environment, visual impressions, attention, listening behavior like active exploration, and many other factors. The same sound pressure at the ears may lead to a different perception if it occurs in a different context.

Dynamic binaural synthesis takes the motion of sound sources, reflecting objects and the listener into account. The headphone signals are adapted in correspondence to the occurring movement. This is necessary to keep the illusion vivid if, for example, the listener moves the head or changes the listening position. There is an increasing demand for the option to explore the entirely or partially virtual scene interactively in **6-degrees-of-freedom (6DOF)**. This means the user can freely move the head in arbitrary directions and change his or her listening position to other positions in the 3-dimensional space.

For the creation of convincing imitations of sound sources in rooms, it is essential to understand sound propagation in actual rooms.

1.2. Sound propagation in rooms

Sound sources emit sound either with a specific directivity or omnidirectionally. In the case of an omnidirectional source, sound propagates in all directions with equal energy. The speed of sound propagation depends on several physical influences like temperature or humidity. According to standard DIN ISO 9613-2 [5], the speed of sound c can be calculated with equation 1.1.

$$c = 343.2 \sqrt{\frac{273.15 + T_{\circ C}}{293.15}} \frac{m}{s} \tag{1.1}$$

For a temperature of 20° C, the speed of sound is around 343.2 $\frac{m}{s}$ [6, p. 3]. In an enclosure like a room, the emitted sound will be reflected, diffracted, and sound energy will partly be absorbed. Fig. 1.1 shows the basic principle of sound propagation in a room.

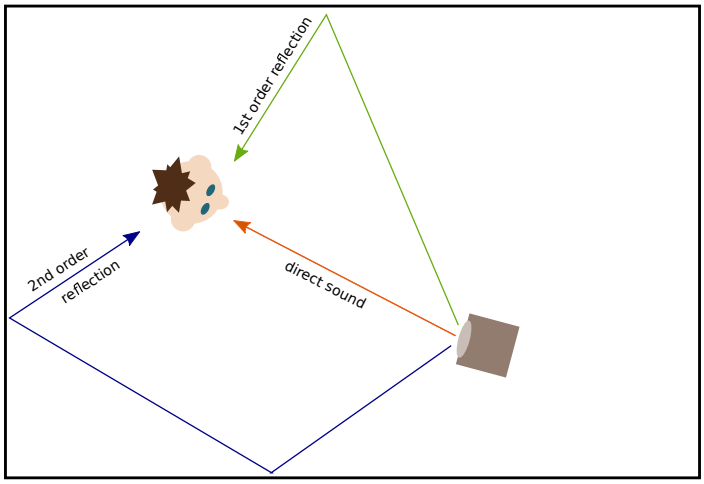


Figure 1.1.: Illustration of sound propagation in a room with direct sound, first and second order reflection arriving at the listener.

In sound propagation, three different phases are distinguished. The first sound arriving at the receiver position is called **Direct Sound**, because it travels the direct path from the sound source to the receiver.

It is followed by the **Early Reflections** arriving from walls and obstacles in different directions and distances, leading to a specific spatial and temporal pattern of sound incidence at the receiver. Fig. 1.1 illustrates examples of the first order and second order reflections. The time duration between the direct sound and the first prominent reflection is known as the *Initial Time Delay Gap (ITDG)*. The acoustic properties of early reflections depend on the geometry and material of the reflecting surfaces.

The early reflections are followed by the **Late Reverberation**. A united and widely accepted definition of the criteria for the transition point between early reflections and late reverberation is still pending. Currently, some investigations and theories focus on physical criteria, and others focus on perceptual properties.

In an ideal diffuse sound field, the energy density is equal for all positions, and sound incidence occurs at equal probability from all directions and with random phase [6]. In the case of such uniform sound propagation, the sound field is called *isotropic*. However, in practice, an ideal diffuse sound field is hardly achieved. According to Jeong [7], it is still not clear whether reverberation chambers produce "sufficiently diffuse" reverberation, although they are constructed with this goal.

If the dimensions of a room are large enough to achieve a sufficient density of modes, it is common practice to assume an ideal diffuse sound field [8], which allows for statistic descriptions. For smaller rooms, a separate consideration of the low-frequency range becomes necessary.

1.3. Special properties of small rooms

With decreasing size of a room, the occurring enclosure effects increase and become significant [6]. If the dimensions of an enclosure become small in relation to the wavelength, the wave behavior of sound has to be considered. Individual resonances in the low frequencies are increasingly audible and can substantially affect the perceived timbre. Small rooms usually have a volume in the range of a few cubic meters to a few hundred cubic meters, like offices, living rooms, seminar rooms, or common anechoic chambers. The standard DIN EN ISO 3382-2 [9] for measuring the reverberation time in ordinary rooms defines rooms with a volume of more than 300 m^3 as *large*. In a small room, for example, resonances can be explored by humming in tiled bathrooms because there they are only weakly damped [10]. The transition from low frequencies, where enclosure effects have to be considered, to high frequencies, for which the room can be regarded as large, is fluent and varies with the room. Based on statistical considerations, Schroeder and Kuttruff [11] developed the equation 1.2 to estimate a transition frequency depending on the volume and reverberation time of a room. This is known as the Schroeder frequency.

$$f_S = 2000\sqrt{\frac{T_{60}}{V}} \tag{1.2}$$

In large rooms, the Schroeder frequency is usually below 50 Hz. Instead, for example, for a seminar room with a size of 150 m^3 and a reverberation time T_{60} = 1 s, f_S = 163 Hz. Rearranging eq. 1.2 leads to eq. 1.3, which allows estimating whether a room should be considered *small* for a given frequency.

$$V > \left(\frac{2000}{f}\right)^2 T \tag{1.3}$$

In addition to these special properties in the low frequencies, small rooms also exhibit other characteristic acoustic properties. For example, early reflections typically arrive at the receiver within a few milliseconds after the direct sound in a dense temporal order. The reflection density or echo density rises more quickly than in large rooms.

Kleiner and Tichy [6] give a detailed overview of the special acoustic conditions in small rooms and how they may affect perception. This section provides only a short summary.

In small rooms, the distances between sources and receiver are relatively small. The listener is likelier to be close to a source or an obstacle that influences the sound propagation. Therefore, assuming a point-like shape of the sound source is not always suitable.

Moreover, in small rooms, different types of sound sources are common. While concert hall acoustics usually consider sound sources like an orchestra or singers, for an evaluation of rooms, e.g., intended for stereo reproduction like studio control rooms, not only instruments of an orchestra (classical music) are of interest but also close-mic recordings, synthesized sounds and percussive sounds [12].

1.4. Listener motion in (small) rooms

This section is based on the section "Relevance of cues from self-motion" in the journal article "Perceptual matching of room acoustics for auditory augmented reality in small rooms" that I published together with Christian Schneiderwind and Florian Klein [311].

When a listener actively changes the own position or (head-) orientation, certain expectations arise for the change of the sound reaching the ears in correspondence to this motion. These expectations are based on the experiences from everyday listening. To understand how to satisfy

these expectations, it is necessary to understand the role of information from the human sense of motion and self-motion in spatial hearing. The vestibular system is one of the contributors to the conscious sensation and guidance of motion and posture. Another contributor is the proprioceptive system. Proprioception refers to the sense of self-motion based on sensory-motor information [13]. So far, only selected aspects of their role in the perception of spatial sound have been examined. Binaural technology is about to become a valuable tool for investigations in this field.

It is known that head motion facilitates sound source localization and improves localization accuracy [14, 15, 16] and helps to resolve front-back ambiguities [17]. Kim et al. [18] observed that listeners move their heads over a wider range when judging source width and listener envelopment than for sound source localization. For the evaluation of timbre, the range of head rotation in azimuth was very low compared to the other tasks. For active changes in the elevation angle, the differences were relatively small and primarily not significant. Active head rotation is also known to improve externalization in a dynamic binaural reproduction [19]. Hendrickx et al. [20] reported that this improvement in externalization persists even after dynamic cues were omitted.

Kondo et al. [21] investigated the influence of active head motion on auditory scene analysis. When listening to a complex scene, the incoming flow of acoustic information is organized in streams that are not instantaneous but built up over time. The organization of streams can also be reset if sudden changes occur in the scene. When moving the head in a stationary scene, the acoustic stream also changes, but the listener should understand that the scene does not change. Kondo et al. [21] observed that with the onset of the motion, the organization of streams was partially reset and reorganized under the consideration of the spatial cues provided by the movement.

Wallmeier and Wiegrebe [22] showed that vestibular and proprioceptional information provides helpful cues for the human echo-acoustic orientation. The observation seems to be based on the same mechanisms as those reported by Kondo et al. [21].

Active self-motion and exploration behavior give access to different cues that the auditory system could use. On the one hand, a listener gains additional information from positional disparity. On the other hand, dynamic auditory cues about the current change of the sound reaching the ears are available.

Listening to a scene subsequently from different positions and perspectives provides more spatial information about it. The brain can collect this information and interpret it. Especially for sound sources in the front, without movement, often in-head localization occurs. Turning the head provides additional information, and after turning it back to the original position, the same ear signals may be interpreted differently considering the collected information.

On the other hand, during motion, the auditory system could use dynamic cues like the current change of the sound level or the recent change of DRR. For example, the acoustic τ (time-to-contact) addresses the current change of sound intensity during a motion of the listener or the sound source [23, 24]. If the listener moves, dynamic auditory cues are available that can be analyzed in combination with proprioceptional cues related to the own motion. In this case, the listener will expect that a certain motion is connected to a specific change in the sound field. This might also be a result of long-term listening experiences. However, this is still more of a hypothesis. So far, studies in this field are rare. The role of the listener's active self-motion in creating auditory illusions is discussed in more depth by Brandenburg et al. [145, pp. 637-645].

1.5. Spatial auditory illusions and their quality

This section is based on the "Introduction" of the journal article "Perceptual matching of room acoustics for auditory augmented reality in small rooms" that I published together with Christian Schneiderwind and Florian Klein [311].

Spatial auditory illusions created with dynamic binaural synthesis are of interest for the realization of Virtual Reality (VR), Augmented Reality (AR), or Mixed Reality (MR). The definitions of these terms vary among the literature [25]. According to Novo [26], VR describes entirely virtual environments which do not correspond to the user's actual environment. In 1994, Milgram et al. [27, p. 3] defined MR as an environment "in which real world and virtual world objects are presented together within a single display" in the context of visual displays. Furthermore, they proposed the famous concept of the Reality-Virtuality continuum with MR covering the whole range between the two extrema of environments consisting only of real elements and fully virtual environments. Therein, AR is defined as a subset of MR where virtual content is added to the natural environment. This is contrary to Augmented Virtuality, a subset of MR, where real-world objects are integrated with the virtual environment. In 2021, Skarbez et al. [28] revised the continuum and the related definitions and dimensions, arguing that the continuum is not entirely continuous, that perfect virtual reality cannot be reached, and that MR embraces a broader range of applications than suggested by the original continuum.

In addition to VR/AR/MR, the concept of **Personalized Auditory Reality (PARty)** was introduced [313] as a future vision. This concept addresses the idea of providing the listener with full control to personalize the perception of the real acoustic environment according to their individual desires. The discussions in this thesis will focus on VR and AR, but the results will also be helpful for the realization of MR, XR, and PARty.

In classic stereo headphone reproduction, auditory images of sound sources are perceived in the head. This phenomenon is called *in-head-localization*. A high-quality binaural audio reproduction is expected to create virtual sound sources localized externally at a meaningful distance from the head. Generating a stable externalization of sound sources in binaural reproduction for every listener remains challenging. Although several influencing factors could be identified already, the related human auditory processing is not fully understood.

In VR/AR/MR, virtual sound sources can be realized with or without correspondence to a visual object. If a visible object can be identified as the sound source, the audio representation needs to match the visual properties like direction, distance, or size. Noticeable **audiovisual mismatches must be avoided** in the realization. Furthermore, the quality of an illusion will be affected if the reproduced room acoustics are too different from the expectations of the listener.

Generally, the virtual elements of an acoustic scene should **meet the listener's expectations** in order to induce convincing spatial auditory illusions. The expectations arise from everyday listening experiences. Observations in several experiments indicate that these expectations differ from person to person and fluctuate even for a single listener over time and within the context of listening. The expectations do not necessarily refer to the real version of the sound field. One might call these expectations "wrong." However, there may be a certain range of variations that will still satisfy the listener's expectations. In many cases, listeners are not even exactly aware of their expectations but still can tell that specific effects are not in line with it.

Understanding the formation of listeners' expectations in their complex, diffuse, and individual nature is one of the main challenges in creating convincing spatial auditory illusions. However, it also bears great potential in increasing the efficiency of implementations based on psychoacoustic optimization. Such optimization includes physical simplifications and approximations, which still create auditory illusions of the desired quality. For evaluating the quality of spatial auditory illusions, the following two constructs have been established.

Authenticity is the stricter quality criterion. According to the definition by Blauert [1], it measures the agreement with an external reference. This means listeners cannot distinguish the virtual sound objects from the corresponding real version in a direct comparison. Brinkmann et al. [29] showed that it is possible to realize a binaural reproduction that cannot be distinguished from the real version of the sound field by the majority of the listeners in the corresponding experiment. However, in this experiment, only interactive head rotation was considered, and authenticity was only achieved for the speech signal. When pink noise was used, most listeners could detect the binaural reproduction.

According to Kuhn-Rahloff [30], **Plausibility** refers to the perceived agreement of an auditory scene with an internal reference. The internal reference corresponds to the expectations of the listener.

For many applications, providing plausible auditory illusions is sufficient. Authenticity is necessary for auralizations intended for the judgment of detailed acoustical properties. Although the distinction between authenticity and plausibility is clear in definitions, for example, for augmented reality, the question arises which quality is required. The goal is to add virtual sound objects to the real environment. Real acoustics take on the role of an external reference. Still, it is not the same requirement as not being able to hear differences to a real version of the virtual element in a direct comparison. At this point, the definitions and the separation of the terms need to be improved. Furthermore, the question of suitable test methods arises.

Besides externalization of the sound sources and audiovisual congruence, **many other factors** contribute to the emergence of plausible or authentic illusions. For a constructive improvement of the system in case of dissatisfying results, it is helpful to identify the limiting factors. An expert group has set up a **Spatial Audio Quality Inventory (SAQI)** [31] to facilitate the perceptual evaluation. It covers the majority of possible quality issues in spatial audio reproduction. Related test methods will be discussed later in the text.

Evaluating the perception and quality of auditory illusions is interesting in this thesis due to its interaction with the auralization of the room acoustics. Reverberation is relevant for immersion, plausibility, externalization, and auditory distance perception. Therefore, it is possible that quality features of VAEs can serve as indirect measures for the perception of room acoustics [32].

Katz and Nicol [33] emphasize that assessing the quality of binaural audio means identifying the perceptual dimensions by which the listeners judge the experience. Quality is multidimensional, and timbre and sound localization have been identified as two main categories for evaluating binaural audio. Their book chapter reviews methods to assess the quality of binaural reproduction.

1.6. Practical relevance of position-dynamic binaural audio

Position-dynamic binaural audio is interesting for virtual acoustic environments and any mixture of real and virtual elements like AR/MR. A prominent field of application is the entertainment and gaming industry. The user is either transferred to a virtual environment and can explore it in various degrees of interactivity, or virtual content like a gaming character is added to the real environment.

In their review on Auditory Augmented Reality (AAR), Yang et al. [34] identify seven main application areas. VR technology is similarly interesting in most of these areas.

- **Navigation and Location-Awareness Assistance** - Spatial auditory cues can be a very natural indicator of directions or object location, particularly outside the human field of view.
- **Augmented Environmental Perception** - spatialized descriptions of the environment can be beneficial for visually impaired people or, in other cases, of limited sight.

- **Presentation and Display** - An application could be the presentation of manuals or virtual instructions for using a given device or for sports exercises.
- **Entertainment and Recreation** - This includes exhibition scenes in the museum, cultural heritage displays, or archaeological sites. Furthermore, besides mobile games, storytelling with AR/MR audiobooks is a popular application.
- **Telepresence Applications** - An augmentation of abstract or rather realistic representations of social characters or actual other people is interesting for collaborative tasks and social interactions over the distance.
- **Education and Training** - VR/AR/MR systems are capable of presenting information vividly and more intuitively, which makes them interesting for training and education purposes.
- **Healthcare** - An example scenario would be the simulation of natural environments to increase physical and mental well-being. Bringing natural outdoor scenarios to indoor listening situations could have restoring effects and be helpful with depression. Furthermore, it seems reasonable to explore therapy approaches for phobias and anxiety disorders, for example, by confrontation with a virtual anxiety scenario.

Moreover, the creation of spatial auditory illusions is very useful in the development of virtual prototypes, whose actual realization would be very expensive, like vehicles, production lines, or buildings [35], or to develop acoustics concepts for not yet constructed rooms. Product sounds or room acoustics can be evaluated regarding their quality under different test conditions [36].

Binaural technology also has become a valuable tool in audiology. According to Stecker [37], traditional laboratory-based testing is still the "gold standard" for clinical auditory assessment and basic research. However, lab-based tasks can be limited in their real-world validity. Listeners between booth walls and desktop displays are provided with inconsistent multisensory context for the auditory stimuli. Already providing 360° visual environments over Head-mounted displays (HMD) was beneficial in speech-based communication and suggested that VR audiological testing of speech-in-noise is feasible in clinical practice [38]. Wearing HMDs while listening to a sound field created around the listener, for example, with loudspeakers, leads to localization errors [39]. VR or AR systems can improve multisensory consistency, for example, by increasing the capability to create spatially coherent audiovisual stimuli with binaural technology over headphones [40, 41]. The additional integration of low-latency tracking methods enables a more natural exploration of auditory stimuli with interactive self-motion. A further advantage is that participants usually perceived the test situation with interactive VR/AR systems as more interesting and rewarding than traditional test setups [37]. The consideration of motion is essential in audiology because it better represents listening in real-life conditions, and types of hearing aids that make use of direction-dependent filtering have to be tested for their ability to quickly adjust to any changes [42].

Besides research for hearing aid technology, dynamic binaural reproduction can be helpful in investigating the human auditory sense. It provides better control of the sound arriving at the eardrums and allows for the creation of manipulated, unnatural sound fields, which is interesting to isolate selected parameters of interest, like listening with different HRTFs [43]. An audio reproduction that reacts dynamically to the listener's motion provides the opportunity to create artificial sound fields by manipulating the sound pressure progress at the eardrums during self-motion. This allows for other test approaches, which are not possible when relying on purely natural sound fields, also in correspondence to proprioception cues.

Binaural auralizations also allow for simulated versions of the sound field with idealized conditions like silent, empty shoe-box-shaped rooms with customizable reflections behavior.

In the research of sound field synthesis with loudspeaker arrays, binaural auralizations can be helpful to compare, for example, different array configurations [44] or room conditions [45, 46] to assess their perceptual effects. So far, perceptual investigations of WFS or Ambisonics have not

been realized with listener translation. However, in the case of WFS, it would be interesting in particular because the sound field is created for a whole listening area rather than a sweet spot.

This thesis considers a test scenario of a virtual loudspeaker standing in a reflecting room environment. It will be auralized over headphones with position-dynamic binaural synthesis for a translating listener. This is a fundamental step for efficiently realizing a virtual loudspeaker setup to be explored by walking around with headphones.

1.7. Objectives of this investigation and Research question

This thesis aims at gaining a better understanding of the perceptual requirements for position-dynamic binaural auralization of room acoustics and identifying potentials for psychoacoustic optimization. As illustrated in fig. 1.2, this demands an assessment of the interaction between the listener's translational motion and the expectations these movements raise about the changes of the auditory perspective and the perception of position-dependent details of the room acoustics.

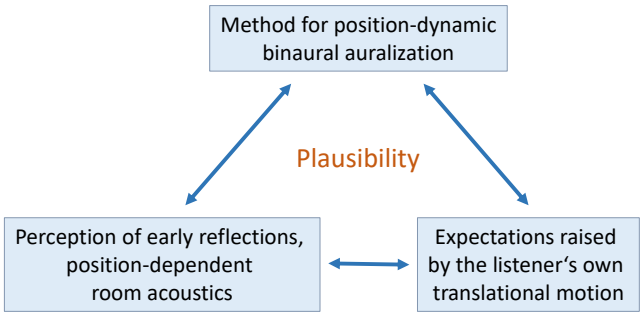


Figure 1.2.: This thesis investigates the interdependencies between the position-dependent room perception and position-dynamic binaural room auralization to be explored by interactive translational movements of the listener.

a) Realize a setup for position-dynamic binaural auralization of rooms

The first step in this thesis is implementing a system for position dynamic binaural auralization of room acoustics. Consequently, the first question to be investigated is:

- **Is the implemented system capable of providing plausible illusions of walking towards, past, or away from the sound source?**

b) Test methods for perceptual evaluation in interactive exploration scenarios

For the majority of applications, the goal is to create plausible auditory illusions. In exceptional cases, for example, for hearing research or very accurate acoustic prototyping, authenticity is desired or at least will be desired, in the future.

At the beginning of this work, there were very few experiments considering interactive position changes. Approved test methods for the perceptual evaluation of binaural auralization for interactive listener translation were lacking.

Consequently, this work is also concerned with realizing a suitable reference in perceptual evaluations of position-dynamic binaural auralizations.

In order to estimate whether the created auditory illusions are plausible, an adequate method to assess plausibility for position dynamic binaural auralizations is required. Therefore, the second research question is:

- **What is a suitable method to evaluate the plausibility of walking towards, past, or away from a virtual sound source?**

On the one hand, this is of interest for examining potential references. On the other, it is also important to find suitable approaches for assessing perceptual deviations of a test condition with a given reference or another test case.

c) Investigate simplified acoustic room representations in position-dynamic binaural auralizations and their effect on plausibility

To explore the potential of simplifying room acoustic representation in dynamic headphone reproduction, systematic simplifications will be realized and perceptually evaluated in psychoacoustic experiments. This is not only of interest to realize position-dynamic binaural auralizations and improve the rendering efficiency. Related investigations will also improve our understanding of the human perception of room acoustics and particularly of early reflections. When the work on this thesis started, the sensitivity to changes in the spatiotemporal pattern was still relatively unexplored by research. Therefore this thesis also addresses questions like the following:

What are the audible effects of early reflections? Are humans able to perceive position-dependent changes in the spatiotemporal reflection pattern? Do humans draw any information from it? Which role does it play in the understanding of the scene? Which audible information gets lost if early reflections are not rendered correctly? Are simplified representations of the early reflection patterns and their position-dependent changes suitable for position dynamic binaural reproduction?

The main research question of this thesis is:

- **Which psychoacoustic requirements can be defined for the accuracy of imitating the position-dependent room acoustic properties in a position dynamic binaural auralization with interactive listener translation to create plausible auditory illusions of walking towards, past, and away from a sound source in a room?**

Answering this question will also demand a physical analysis of room acoustic differences between different listening positions. This thesis aims at understanding the relationship between perceptual effects and observed physical properties.

Rather than attempting to propose any specific algorithms for position dynamic binaural auralizations, this thesis focuses on exploring the perceptual requirements and boundaries for simplified or impaired acoustic room representations concerning the plausibility of the created spatial auditory illusion.

1.8. Outline of the thesis

This thesis explores the perceptual consequences of impaired and simplified room representations in position dynamic binaural room auralizations over headphones, particularly their impact on the plausibility of spatial auditory illusions. For example, if a listener moves within the sound field of a stationary sound source, the direct sound will change with the listening position. In addition, also the local acoustic conditions change, for example, the spatiotemporal pattern of early reflections or parts of the late reverberation that are not perfectly diffuse.

After this introduction, the fundamentals of spatial hearing, room acoustics and binaural , room perception, and listener translation are summarized in Chapter 2. Furthermore, this chapter provides a state-of-the-art overview of approaches to realize position-dynamic binaural synthesis and methods to evaluate the plausibility of spatial auditory illusions.

Chapter 3 describes the implementation of the binaural reproduction system used within this thesis. The technical properties of the related devices, like tracking modules or headphones, are presented. Besides the developed rendering tool pyBinSim, adequate BRIR data sets for listener translation are required. The chapter documents the creation of such data sets based on acoustic measurements with a head-and-torso-simulator and room acoustic simulations. Furthermore, an informal attempt to evaluate the plausibility of the first binaural walk-through experiences with the new system is described.

Chapter 4 presents the experiments conducted within this thesis. The first two experiments, documented in sections 4.1 and 4.2, explore the impact of various simplifications of the original measured BRIR sets measured for walking towards (and away from) a virtual sound source in a pretty dry listening laboratory and a more reverberant seminar room. In both studies, the listeners wore an HTC Vive head-mounted display showing a neutral grid environment. Consequently, the participants did not get a visual impression of the auralized environment. The evaluation was mainly based on the interaction of their own (walking-)motion and the perceived changes in the dynamic binaural audio. The results of both experiments reveal the considerable potential for simplification but also highlight some boundaries.

The first experiments were realized based on a uniform position grid of BRIR measurements of simulations with a resolution of 25 cm. The experiment described in section 4.3 evaluates the perceived continuity of the auralized sound field to determine the minimum BRIR grid resolution for listener translation with different source signals.

The first two experiments evaluated plausibility in a single-scene (equivalent to single-stimulus) test design concerning a virtual reference realized with the initially measured data set. The investigation presented in section 4.4 evaluates the plausibility with and without the availability of a real counterpart as a hidden reference.

The final experiment in this thesis, presented in section 4.5, uses an approach to evaluate plausibility concerning a virtual reference that cannot be assumed to be perfect. With this setup, the study addresses various impairments and simplifications representative of state-of-the-art auralization methods that interpolate or extrapolate from sparsely measured room information.

Chapter 5 discusses observations from the row of investigations in relation to other studies in this field. It ponders which conclusions can be drawn from the experimental results gained within this work and to which extent these can be generalized to other rooms, listening scenarios, and different auralization methods. Section 5.2 provides a summary of the studies and findings in this thesis, and section 5.3 formulates a list of conclusions. It will also point out the limitations of the presented studies and give an overview of the remaining open questions. The final section suggests consequent future work.

2. Fundamentals, State of the art and Literature Study

This chapter contains several text passages from the journal article "Perceptual matching of room acoustics for auditory augmented reality in small rooms" that I published with Christian Schneiderwind and Florian Klein [311].

This chapter provides the definitions of relevant terms and explains the fundamental principles the investigations in this thesis are built on. Three main areas are covered in this chapter.

First, the basic concepts and properties of acoustics in real rooms are summarized. Relevant psychoacoustic phenomena will be explained. The second section describes the process of creating spatial auditory illusions with (dynamic) headphone-based reproduction. An overview of typical signal processing concepts and fundamental psychoacoustic requirements for dynamic binaural synthesis is given.

The last section in this chapter rolls out the known issues of transferring the acoustics of real rooms to a headphones-based dynamic auralization. Particular attention will be paid to the acoustical differences at different positions in a room. A detailed understanding of the physical deviations of the acoustics within a room and their perception are fundamental for an efficient realization of a virtual representation that can be explored by interactive head rotation and translational motion.

2.1. Spatial hearing

Spatial hearing describes the relations of an auditory event and correlated other events, like sound events or physiological processes. The auditory event is defined by what the listener perceives and how the event appears to the listener, for example, the apparent location, apparent width, and shape of a source image or its loudness. Fig. 2.1 shows the head-related coordinate system. Directions in the horizontal plane are described by the azimuth angle, up and down directions are addressed by the elevation angle.

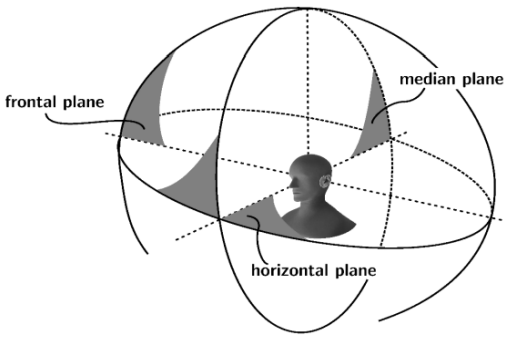


Figure 2.1.: Head-related coordinate system, taken from [47].

Estimation of direction The first lines of this thesis (sec. 1) explained that the human auditory sense analyses the differences of the sound arriving at the left and the right ear. Interaural time differences (ITD) and interaural level differences (ILD) are used to estimate the direction of the sound

source. At lower frequencies of up to about 1.5 kHz, ITDs are the dominant cues for localization. Above 4 to 5 kHz, ILD becomes dominant. In the frequency range in between, both contribute to localization [1]. In addition, spectral cues play an important role in the localization of sound sources. Particularly higher frequencies above about 5 kHz cues introduced by the shape of the ears are helpful [4]. These cues are encoded in direction dependent head-related transfer functions (HRTF). Moreover, small head movements can help to resolve front-back confusion or other effects related to the cone of confusion, for example, for localizing elevated sound sources. If a listener can only use one ear, sound localization is still possible to some extend with the help of monaural cues. These cue are mostly based on direct-dependent spectral characteristics. Monaural are also essential for localization in the median plane.

Auditory perception of egocentric distance and distance change The auditory distance perception in real and virtual acoustic environments has been studied intensely throughout the last decades. Zahorik [48] provides a summary of the results. After that, the essential cues for auditory distance perception are loudness, the direct-to-reverberant ratio (DRR), spectral cues, and binaural differences. Further aspects like familiarity can play a role as well.
For sound sources closer than about 1.6 m, listeners tend to overestimate the physical distance, and for farther sources, usually, an underestimation occurs. This indicates that the perceived distance can differ from the physical distance. Based on a thorough analysis of experimental data, Zahorik proposes a description of the relationship between perceived distance r' and the physical distance r

$$r' = \kappa r^a \tag{2.1}$$

with κ being a constant number of about 1.32 and a being influenced by various factors like the person or the experimental condition. a shows substantial variations around an approximate average of 0.4. Xie [4, p. 8] provides a nice summary.

If visual information on the source distance is available, it dominates the audio-visual estimation of the egocentric distance to the corresponding object [49].

Nielsen [50] observed that in an anechoic chamber, the auditory distance perception of the participants did not correspond to the physical distance. In contrast, in all reverberant rooms in the study, the distance could be estimated quite accurately.

So far, meager attention has been paid to the perception of continuous distance changes as well as its interaction with physical self-translation.

Perception of distance change due to walking Especially when walking toward the source, listeners have certain expectations from their everyday experiences. A listener typically knows how the auditory impression of a loudspeaker in a room changes while walking through the room. In a virtual acoustic environment, listeners will compare the sound reaching their ears to an inner reference.

According to Kuhn-Rahloff [30], plausibility describes the agreement of the hearing with the listener's internal reference. Thus, it will serve as a measure of the agreement of the heard distance change with the expected one.

2.2. Acoustic impulse responses and room acoustic measurements

A linear time-invariant (LTI) system can be characterized by the response at the output to a unit impulse signal at the input. The output signal is called *impulse response*. The unit impulse signal is an idealized mathematical construct which cannot be achieved in practice.

Furthermore, the acoustics within a real room are usually not precisely time-invariant. Slight changes in temperature or humidity, as well as air movement due to wind or, for example, moving people, can cause time variances. However, for the investigation and the mathematical analysis of room acoustics, it is helpful to assume an LTI system. Typically the following kinds of impulse responses are used to characterize the acoustical behavior of a room.

Head-related impulse responses (HRIR) describe the filter effects caused by the shape of a person's head, torso, and ears. An HRIR characterizes the transfer path from a sound source to the two ears of a listener or dummy head in the free field. In practice, HRIR measurements are usually conducted in anechoic rooms, which are never perfectly anechoic.

Room impulse response (RIR) describes the transfer characteristics for sound traveling from the source through the room to the usually omnidirectional receiver. For a standardized measurement procedure for RIRs, further demands must be met [51, 9].

Binaural room impulse response (BRIR) characterizes the transfer path from a sound source to two ears of a listener or dummy head. BRIRs contain spatial information about the acoustic scene, like the directivity, position, and orientation in the room and relative to the listener, the listener's head-(and-torso-)related spatial information, or a similar substitute. Moreover, it contains room acoustic cues like the Spatio-temporal pattern of early reflections representing the geometric arrangement and the acoustic behavior of the different surfaces like reflection, scattering, and diffraction.

Spatial room impulse responses (SRIRs) are determined for a specific number of receivers that form an array by being located in a certain geometric arrangement. Spherical microphone arrays are common for acoustic measurements for direction-dependent capturing of the sound field created by a single sound source in a room. This allows for signal processing and analysis in the spherical harmonics domain. Furthermore, several distributed spherical arrays can form a unit that may be analyzed altogether.

Measurement of acoustic impulse responses - in the room

Within this work, the swept sine method [52] was used to measure acoustic impulse responses and characterize transfer paths in a room. This method was chosen because it is very robust with regard to minor time variances in the transfer path as well as to mismatches between the sampling clock of the signal generation and recording. The method is based on the idea of using an exponential sine-sweep, which allows for the simultaneous extraction of the linear impulse response of the system and the impulse responses for each harmonic distortion.

2.3. Virtualization of room acoustics using binaural technology

2.3.1. Creating spatial auditory illusions over headphones: Binaural Technology

Section 1.1 described the fundamental principle of binaural synthesis. The basic idea behind binaural technology is to mimic the acoustic cues necessary for spatial hearing at both eardrums. This can be achieved by convolving a dry mono audio signal with binaural filters like an HRTF. However, this thesis is concerned with the binaural auralization of rooms. To consider room acoustic information in spatial auralization, HRTFs are replaced with BRIRs. Fig. 2.2 visualizes the basic principle.

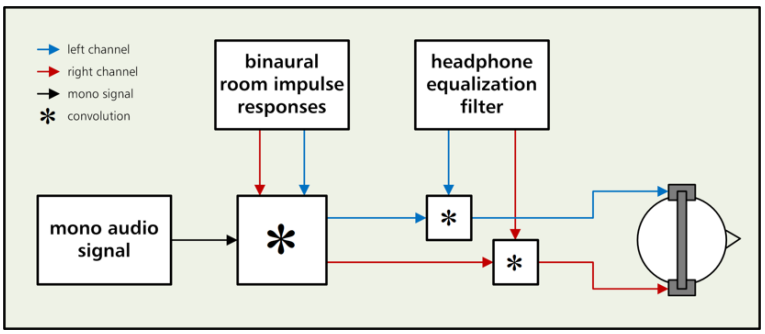


Figure 2.2.: Basic principle of binaural auralization.

Headphone compensation Spectral cues play an important role, for example, in sound source localization. Hence, systems for binaural auralization need to preserve the spectral information in the best possible way. Therefore, influences or disturbances caused, for example, by headphones should be compensated. The headphone response HpIR or the headphones transfer function HpTF characterize the transmission from a headphone to the listener's eardrum. Furthermore, acoustic interaction occurs between the headphones and the external ear [4, p. 265]. The fundamental idea behind headphone compensation is to determine the HpIR, for example by repeated measurements. To compensate for that influence, an inverse filter could be calculated and applied in addition. For practical reasons, a direct inversion of the filter is not desirable. For example, it is of interest to avoid strong amplifications in the very low and very high frequencies. As an alternative, Schärer and Lindau [53] proposed different methods to calculate filters for headphone compensation, for example, based on the least-squares criterion. Frequency-dependent regularization is employed to minimize the error between a target function and the equalized frequency response. This method was used for the realizations within this thesis.

The hpir can be measured for the individual listener or, for example, with a head-and-torso-simulator. Lindau and Brinkmann [54] showed that even for non-individual binaural recordings, an individualized headphone compensation will lead to better perceptual results.

Individualization Each individual has a different size and shape of head and torso. Therefore, the HRTFs and BRIRs vary slightly for each person. Listening with the HRTFs measured at another person or a dummy head can affect the localization accuracy and increase errors like front-back-confusion [55]. Furthermore, a mismatch in the ITD of the HRTF or BRIR was noticed in previous studies [56]. If the provided ITD is smaller than the required one, the sound source will move along when turning the head. If the ITD is larger, the source will move in the opposite direction during head rotations. Therefore, they suggested conducting at least a personalization of the ITD.

A perceptual evaluation of recording realized with several common dummy heads, showed that all of them caused slight deviation in localization [57]. However, humans can also adapt to different HRTFs and learn to hear with another person's ears [43, 58].

In this thesis, binaural auralizations are realized only based on generic head related cues of the *Kemar 45ba* dummy during measurement or the *Neumann KU100* dummy used to create BRIRs based on room simulation. Therefore, it is worth noting that their use may introduce inaccuracies in localization.

Dynamic binaural synthesis for head rotation and position changes

The goal of dynamic binaural synthesis is to provide spatial auditory illusions that endure listener motion and remain stable in the room. This goal requires that the headphone reproduction is dynamically adjusted to the listener's motion. The motion is captured and the audio reproduction has to respond to any changes in the scene without a noticeable delay. Psychoacoustic studies

found that a delay becomes noticeable from about 58 ms [59] to 100 ms [60]. Wenzel [61] reported similar delays depending the speed of the source motion with 92 ms for slow motion and 59 ms for fast movement. To ensure a short system response to movements of the listener, the system has to perform an ongoing convolution of shorter parts of the dry source signal and the binaural filter. This is called partitioned convolution.

Partitioned convolution for dynamic auralization in real-time Partitioned convolution is an efficient method of realizing FIR filtering. Signal and filters are divided into blocks or chunks. If these blocks have the same size throughout the process, the term **uniformly partitioned convolution** is used [62]. Uniformly partitioned convolution can be realized, for example, with the Overlap-Save approach, which is illustrated in fig. 2.3. A K-point sliding window is used to calculate the input FFT. After the multiplication in the frequency-domain, and the K-point IFFT, K-B samples are discarded because of time aliasing. But the resulting signal contains a correctly convolved block which can be used right away without further additions.

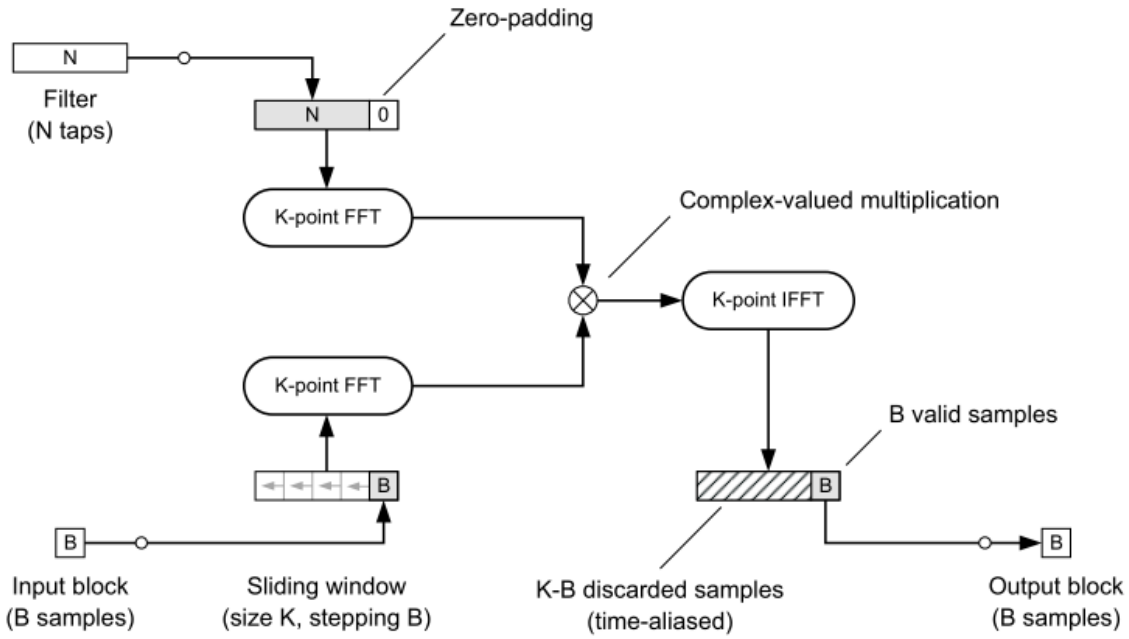


Figure 2.3.: Overlap-Save approach for partitioned convolution [62].

In **non-uniformly partitioned convolution** the signal and the filter are split into chunks of different sizes. For example, keeping the block size small at the beginning of the BRIR filters can be effectively reduce the response times of a dynamic binaural audio system to movements of the listener. However, working with varying block sizes requires accurate synchronization.
A beneficial way of applying this concept to dynamic binaural synthesis is to use short blocks for the early part, in order to provide a quick system response. Assuming the later part of the binaural room impulse is not as critical with regard to a quick update, it is possible to use longer chunks for that part of the impulse response.

2.3.2. Basic system for auditory augmented reality

For auditory augmented reality (AAR), the audio reproduction is synthesized considering information about acoustical properties acquired from the environment. This can be done either a priori, for example, by pre-measurements and offline pre-processing or in real-time "on-the-fly." Over the years, various approaches to synthesizing binaural room impulse responses for a 6DOF listening area have been presented. These are all built on (a small amount of) a priori information about the room and the sound sources within. The available information can be acoustic impulse responses, measured with a single (omnidirectional) microphone [63], a head-and-torso-simulator [64, 65] or microphone array solutions [66, 67, 68, 69, 70, 71]. Besides, for example, semantic and visual information can be used to estimate acoustic properties [72, 73]. BRIR synthesis can be realized either by pure simulation, for example, based on ray-tracing [74, 75], wave-based simulation approaches [76] or delay networks [77, 78] or by manipulation of measured impulse responses, like interpolation [79, 80], extrapolation [81, 64, 82, 63] or shaping of the late reverberation tail [83, 84, 85]. Systems that do not rely on a priori knowledge are desired because their use is not limited to rooms or environments for which the predetermined information is available. Such systems attempt to analyze the listener's current environment in (close to) real-time, based on streamed microphone signals or/and signals from other types of sensors, and adjust the reproduction accordingly. Depending on the complexity of the capturing system and the desired level of detail, the computational effort of the scene analysis can be pretty high. It may not meet (close to) real-time requirements. Sophisticated and quite robust approaches for the blind estimation of the room impulse response [86], reverberation time (RT), or direct-to-reverberant-energy-ratio (DRR) has been proposed [87]. RT or early-to-late-energy-ratios (ELR) can be estimated for broad or selected frequency ranges [88, 89]. This is, for example, used for automatic speech recognition and the necessary dereverberation.

For AAR, a targeted combination of these approaches with established methods for binaural rendering is desired. In order to create an efficient realization capable of adjusting the reproduction in (close to) real time, a psychoacoustic optimization of both scene analysis and spatial audio rendering is inevitable. This demands understanding the contribution of the single physical parameters and the required accuracy, which still generates spatial auditory illusions of the desired quality.

2.4. Perception of room acoustics

The perceptual effects of room acoustics have been studied in various research areas, highlighting the multidisciplinarity [90]. These areas include, for example, speech intelligibility, architectural acoustics, sound reproduction, and echolocation. Room acoustics provide valuable information for spatial hearing, like distance perception and externalization, as well as auditory scene analysis. In contrast, it can also impair source localization or speech understanding. In the context of studying the preferred acoustical properties of concert halls, Vorländer [91] summarizes that "the three most important factors (loudness, reverberance and spatial impression) explain most of the statistical variance when comparing the acoustic conditions in auditoria." He also points out that open questions remain, for example, regarding "the listener's sensitivity to changes in a sound field regarding those subjective aspects."

For quantifying and adjusting the perceived acoustical similarity of two rooms, the correlations between perceptual quality features and physical measures of room acoustics must be understood. These correlations have been subject to research for several decades, especially in the context of concert halls. As a result, various room acoustic parameters determined from the physical properties of the sound fields in such halls have been developed to describe and predict their perception. A selection of such parameters is summarized in standard Int. Organization for Standardization [51]. These include, for example, RT, early decay time (EDT), clarity indices (C80 and D50), sound strength (G), as well as the interaural cross-correlation coefficient (IACC). The just-noticeable difference (JND) is the minimum charge of the parameter, which produces a noticeable variation in the sensory experience [92]. The standard provides concrete values for the JNDs of the listed room acoustic parameters. However, additional studies indicate that JNDs can vary considerably from the specified values and can depend on other conditions like the type of signal, the frequency content, or the absolute values of the parameters of interest [93, 94, 95]. Appendix B provides a more detailed overview about how the various parameters are calculated and what is known about related JNDs so far.

In their review on room acoustical parameters as predictors of room acoustic impression, Weinzierl and Vorländer [96] conclude that after "more than 50 years of research on developing psychoacoustical measuring instruments for the concept of 'room acoustical impression'" and "more than 100 years of research on the development of physical measures which could serve as technical predictors for these perceptual qualities", "the state of the art is surprisingly unsatisfactory." In another review, Bradley [97] shows that it is still unclear how some of these parameters should be calculated best. For example, findings by Barron suggest averaging EDT values from 125 Hz to 2 kHz works best, whereas the ISO standard suggests a mid-frequency average. Bradley also points out that more research on JNDs and their complexity is needed as they are essential to understand the correlations of such measures to the perception.

In addition to the remaining open questions about objective measures for the perception of concert hall acoustics, it is also not sufficiently clear to which extent this knowledge is valid for the acoustics of small rooms. Standard DIN ISO 3382-2 [9] describes a procedure to estimate the reverberation time in standard rooms. No other parameters are listed. Some aspects of the established parameters have been motivated by the properties of the human auditory sense, which is still the same in small rooms. However, physical conditions like typical listening positions concerning the sound source, the types of sound sources, the decay behavior, and the progress of echo density after excitation differ from performance rooms. Are the JNDs the same under these conditions?

Moreover, cognitive effects like becoming familiar with and adjusting to the room can play a role. The auditory room perception can vary, although the physical sound field remains the same [145]. Such effects might even dominate the influence of physical details under certain conditions.

van Dorp Schuitman et al. [98] point out that the auditory perception of room acoustics also depends on the type of the source signal. Hence, there is a general shortcoming in predicting the perception of room acoustics only from parameters estimated from room impulse responses since this

approach does not take the type of signal into account. Instead, the authors propose a new concept of parameters determined from binaural recordings of the sound field in the room. The correlation with the perception of reverberance, clarity, apparent source width, and listener envelopment is better in most cases with this method.

In the context of loudspeaker reproduction in rooms, it is essential to distinguish between the bass-frequency range in which room modes can cause standing wave behavior and the range of mid and high frequencies Toole [99, p. 153-156]. For the frequencies above the transition range, rooms cause changes in the timbral and spatial perception of loudspeaker reproduction. Prominent early reflections can cause audible comb filter effects and shift the perceived position and size of the image of the sound source. Room resonances can cause an audible change in timbre as well. Later arriving reflections contribute to a sense of spaciousness or listener envelopment.

Studying how the acoustics of small rooms influence the perception of a multi-loudspeaker reproduction, Kaplanis et al. [100] observed that the perceptual differences were based on two main dimensions which can be characterized by the four perceptual constructs *reverberance, width & envelopment, proximity, and bass*.

Zahorik [101] studied the perceptual similarity of rooms with 15 small-room auralizations based on measured and simulated BRIRs. He concluded that "when at-the-ear signal levels were held constant, the rooms differed along just two perceptual dimensions: one related to reverberation time (T_{60}) and one related to interaural coherence ($IACC$)" [101, p. 1]. The study did not consider listener motion, and in each room, only one listening position approximately in the center of the room was taken into account.

2.4.1. Perception of a single reflection

"A reflection which is perceived at all does not necessarily reach the consciousness of a listener. At low levels, it manifests itself only by an **increase in loudness** of the total sound signal, by a **change in timbre**, or by an **increase in the apparent size** of the sound source" [10, p. 163].

For a reflection to become noticeable by any kind of audible effect, a certain level is required. This threshold is the **absolute threshold of perception** or **audibility threshold** of a reflection, also referred to as the **reflection masked threshold (RMT)**. It depends on the time delay relative to the direct sound, direction of incidence relative to the listener and relative to the direct sound, and maybe other reflections and the kind of test signal [6, p. 240],[10, p. 161],[102, 12, 103]. For example, for anechoic speech and tone stimuli, Begault et al. [104] reports absolute thresholds (70.7 % level) between 12-31 dB below the level of the direct sound. Adding reverberation with $T_{30} = 0.6s$ raised the threshold by 7 dB. These results were obtained considering delays of 3-30 ms for the reflection relative to the direct sound and several different directions of reflection and direct sound. "Direct sound" and "reflection" were imitated by two loudspeakers reproducing the same signal, with one of them being delayed and attenuated.

Buchholz et al. [102] developed a Room Reflection Masked Model (RMM), which distinguishes separate modules for simultaneous and non-simultaneous masking to achieve a better understanding of the underlying auditory processes. The model is built on a multidimensional function that determines the RMT, considering nine parameters like the direction of direct sound, the direction of test reflection, delay of reflection. Buchholz [105] studied the traditional and simultaneous RMTs for a 200 ms long broadband noise stimulus. At least for this signal, simultaneous masking was dominant for reflection delays below 7-10 ms, and forward masking was dominant for larger delays. Furthermore, **Binaural Masking Level differences (BMLD)** can be examined to quantify the contribution of binaural hearing. BMLDs reflect the difference the thresholds for diotic and dichotic stimulus presentation. For reflection delay of 7-10 ms, BMLDs of about -7 dB were observed which indicates a suppression mechanism. However, for delays beyond 7-10 ms, a positive BMLD of about 3 dB was measured. This suggests that for later reflections, binaural hearing enhances the detection of reflections.

Precedence effect and summing localization If two coherent sounds of a similar level arrive from two different directions with a very small time difference of 0 to about 7 ms, so-called summing localization occurs. The sounds are not perceived separately, but as one fused phantom source. For a delay of 0 ms the phantom source is localized in the middle between both directions of sound incidence. With increasing delay, the apparent source direction moves more and more to the direction of the earlier sound incident.

Starting from a certain delay which varies with the type of signal, the sound image will be localized in the direction of the first incoming wave front. This effect is called the **law of the first wave front**, the precedence effect [106] or the **Haas effect** after [107, 108].

The experiments are based on a lead signal and a delayed and, in many cases, attenuated copy of it, the lag.

With a further increase of the delay, at some point lead and lag are perceived as two separate events which appear at different locations. "The briefest lead-lag delay at which subjects report perceiving 'two sounds' or are able to accurately identify or discriminate the lag location" is called **echo threshold** [109, p. 2]. The echo threshold varies substantially with test stimuli, the test paradigm and the detailed definition of the echo threshold criterion. As a consequence, "a wide range of echo thresholds [...] ranging from 2 to 100 ms and more" was observed [109, p. 2].

Fig. 2.4 visualized the localization depending on the delay of the lag. It indicates that summing localization lasts until a delay of about 0.8 ms.For delays larger than 0.8 ms, up to 30 ms, the precedence effect can be observed. The localization is determined by the direction of the first of the coherent contributions [108]. Contributions or wave fronts after delays longer than about 30 ms are perceived as echoes. As discussed, both threshold values can vary substantially, for example, with the test signal, with the individual listener and with the test paradigm.

The precedence effect has been studied intensely since its discovery. Zurek [110] and Blauert [1] provide reviews of the early studies, Litovsky et al. [111] reviewed investigations until 1999. They describe the precedence effect as a group of phenomena and divide these phenomena into measures of fusion, localization dominance and discrimination suppression. Furthermore, buildup and breakdown effects of the precedence phenomena are discussed. These are addressed later in this thesis. Litovsky et al. [111, p. 1652] conclude that despite many studies on the precedence effect are motivation by the "desire to understand how the auditory system processes multiple arrays of directional cues in reverberant spaces", but "many of the studies do not directly address 'real-world' issues".

Brown et al. [109] reviewed additional results of the following 15 years. The commonly used lead-lag paradigm is a keen simplification of sound propagation in rooms. Brown et al. conclude that "a more ecological understanding of the precedence effect as a mechanism for the preservation of accurate sound localization in reverberant environments [...] will ultimately require more ecological approaches to its study" [109, p. 24]. Few studies considering more than one lag, more than one lag direction, or signals different from click-trains have been addressed in this review.

Moreover, representing a reflection as an ideal impulse is a substantial simplification either. Natural reflections usually underlie a spatial and temporal spread that depends on the directivity of the sound source, the reflection properties of the corresponding surface, and the geometrical constellation of sound, reflector, and receiver. It was shown that these natural reflection properties result in a considerably different appearance of the precedence effect [112, 113] and for surfaces in close distances (< 50 cm) additional near-field effects occur [114].

Barron [115] was interested in exploring other perceptual effects of a single reflection in a concert hall. He conducted an experiment in the lead-lag-paradigm with the lead arriving from the front and the lag arriving from 40° to the left. The study was realized with orchestral music and two participants. The observations and results were visualized as shown in fig. 2.5. This figure can be found in many well-known textbooks on acoustics, for example, in [6, p.], because it provides an excellent overview of the occurring perceptual effects. For lag levels of about -20 dB relative to the lead, the audibility threshold can be observed, with a maximum at a delay of about 10 ms. The

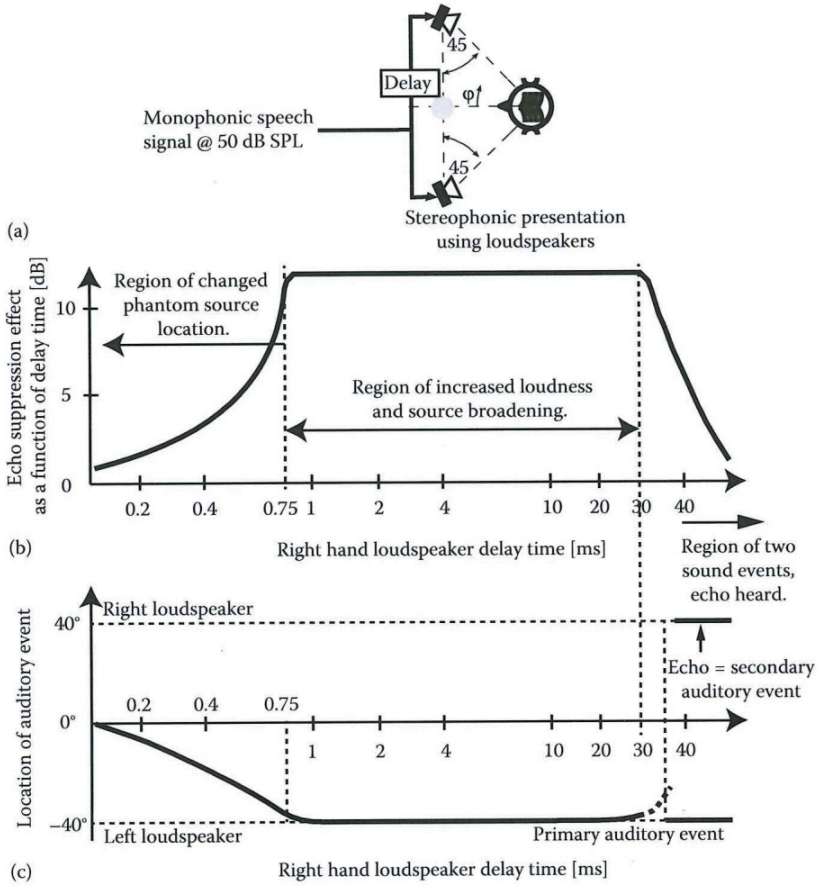


Figure 2.4.: Occurrence of the precedence effect. (a) Loudspeaker setup used in the investigation, (b) Necessary level increase for moving the apparent source position from the left loudspeaker in the right direction, according to Haas (1951,1972), (c) Apparent direction of the phantom source when both sound are reproduced at the same level, according to Madsen (1970). (Taken from [6, p. 253])

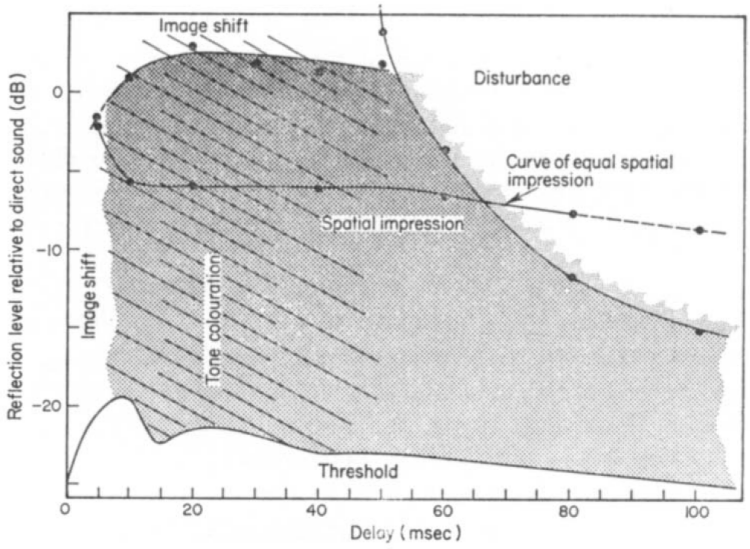


Figure 2.5.: Subjective effects of a single reflections - a visualization created by Barron [115].

indicated "Disturbance" relates to echo effects that occur beyond the echo threshold.

For very short delays of up to about 8 ms, image shift effects have been reported. If the reflection exceeds the energy of the direct sound, the **image shift effect** can still be observed at considerably later delays of the reflections.

In the graph also an area of tone coloration is marked. If, in addition to the direct sound, a coherent signal arrives at the listener with a slight delay, **comb-filter effects** are likely to become audible as coloration, which characterizes the deviations in timbre [116].

Already in 1969, Bilsen and Ritsma [117] reported that "when a sound and its (delayed) repetition are added together and listened to a subjective tone is evoked with a pitch corresponding to the reciprocal value of the delay time." They called this effect repetition pitch. It is well-known that the superposition of coherent signals, like the direct sound and its reflections in the room, leads to comb filter effects [118]. These effects are audible as coloration, which varies with the delay between the original sound and its repetition, as well as their energy relation.

2.4.2. Spatiotemporal structure for early reflections

One of the major questions in the field of position dynamic room auralizations is the role of early reflections and the sensitivity to deviations in their spatiotemporal pattern, as well as the properties of single reflections.

The latest in-depth evaluation of room acoustical simulation tools [75] revealed that the perceptual difference between measurement and simulation are deviations in apparent source position and coloration. According to the authors, these differences can largely be "traced back to the simplified use of random incidence absorption and scattering coefficients and shortcomings in the simulation of early reflections due to the missing or insufficient modeling of diffraction" [75, p. 1].
Also, in the interpolation or extrapolation of BRIRs, considering all the details of the early reflection pattern is challenging. Therefore often, simplifications are applied [79, 80, 69]. For this reason, it is crucial to understand the level of detail required to provide a room acoustic impression without perceptual discrepancies. This is especially interesting for sources and listeners moving in 6DOF because the relative spatiotemporal pattern changes with each position change.

As discussed earlier, adding a single strong reflection to the direct sound will cause comb filter effects. These can lead to audible coloration also in naturally complex structures of early reflections [119, 120, 116]. The character of this effect changes with the delay of the reflection and its individual properties relative to the direct sound. Such effects must be considered if auralizations aim at creating authentic auditory illusions. In these cases, an estimation of the geometrical arrangement of the surroundings is necessary. However, even with the goal of creating a virtual copy of a real sound object, it is not clear whether listeners precisely expect the original progress of timbre during motion.

Adding first-order image source reflections to a rotating directional sound source in a small room can lead to considerable shifts of the apparent source direction [121]. The addition of further reflections did not cause additional localization shifts, and it only smoothed the transition. Similar observations were obtained by Steffens et al. [122]. Early reflections also influence other spatial aspects than the apparent location of the sound source [123]. One example is the apparent source width (ASW) [124] and the apparent sound level of the sound source. The apparent sound level is probably not critical for audio-visual coherence. An interesting question is whether this increase in the apparent sound level can reach the threshold at which it becomes critical for perceptual room matching. For ASW, IACC_{early} is considered a good indicator [125]. However, IACC_{early} varies considerably with the orientation of the head relative to the sound source and can also vary with distance. Thus, maybe a matching of IACC_{early} is only interesting for the person facing the sound source. In addition, ASW varies with the reflection's angle of incidence [126]. More studies considering the diverse properties of natural room environments are required.

Shinn-Cunningham and Ram [127] observed that the sensitivity to differences in the early reflection pattern due to different positions in the room is limited, and so is the understanding of the own listening position in the room. Studies with blind people [128] reveal that long-term training can improve the capability of extracting information about the environment from auditory impressions. Klein et al. [129] found that after a short training, only very few listeners could confidently assign the listening perspective to the corresponding visual perspective of the room if the direct sound is kept constant. Only special cases, like a listening position close to a wall, were recognized reliably by most participants. In the case of a weak direct sound, for example, behind a sound source, the audible differences are most prominent, as a considerable shift of the apparent source location towards the first dominant reflection [130].

Minor differences in the reverberant part could be masked by the strong direct sound as discussed, for example, by Buchholz et al. [102] and Welti and Jensen [131].

A perception-based simplification algorithm was introduced by Hacıhabiboğlu and Murtagh [132], which aims at reducing the number of early reflections needed for the auralization. Based on a prediction model, only image sources that contribute to the perception of the sound field are selected for the auralization. In the perceptual evaluation, the proposed method showed no significant degradation concerning localization performance and perceived spatial quality features such as Presence, Spaciousness, and Envelopment.

Further details in the structure of early reflections have received only little scientific attention so far. For example, considering edge diffraction in the simulation of early reflections has been shown to be audible for selected signals in an ABX test paradigm with monaural auralizations [133, 134]. The latest round-robin comparison showed that room simulations without consideration of edge diffraction still produce plausible auralizations [75]. In AR scenarios, psychoacoustic evaluation of such details in the early reflections like edge diffraction, but also near-field and shadowing effects are still pending.

In summary, it is known that the sensitivity to the physical details in the early reflections and their spatiotemporal pattern of arrival at the listener is limited. However, simplifications in the early reflections can cause noticeable colorations, change the apparent source width and affect correct source localization in direction and distance, which can also affect audio-visual coherence.

Studying the effect of the various physical parameters in the spatiotemporal patterns of the early reflections on perception requires considering its interrelation with the properties of the direct sound and the late reverberation. In addition, the visual impression and the listener's expectations must be considered. The perception of early reflections is a complex field that is not yet understood in detail. This section can only provide an overview and point out that still, more research is necessary to acquire a complete understanding.

Early-to-Late-Energy-Ratios and their relative change

Early-to-Late-Energy-Ratios (ELR) play an essential role in predicting the perceptual quality of concert halls. Examples are the clarity indices C80 and C50 and the direct-to-reverberant-energy-ratio DRR as special cases of ELR. The DRR is known to be an essential cue for the auditory distance perception in rooms and is known to vary with the room [2].

In different rooms, similar DRR values correspond to different source distances. As discussed by Shinn-Cunningham [135] the human auditory sense needs to adapt to the acoustical properties of each room in order to enable an accurate distance estimation. Furthermore, Wendt et al. [136] and Laitinen et al. [137] showed that the variation of directivity influences auditory distance perception. A sound source with a different directivity or orientation will impact the position-dependent progress of the direct sound energy and consequently affect the progress of the DRR. It remains open to how accurately the progress of the DRR has to be imitated to achieve plausibility. This could also be a matter of the spatial, temporal, and spectral distribution of the direct and the reverberant sound

energy.

In concert halls, C80 is used to estimate the perceived clarity of the room acoustic with (orchestral) music, while C50 is associated with speech [10]. Both parameters are interesting because they consider the perceptual fusion effects between direct sound and early reflections that occur in rooms based on the mechanisms of the precedence effect. Furthermore, they address that the time range of this perceptual fusion depends on the type of source signal. For small rooms, the clarity indices are also interesting due to their correspondence to the perceptual mechanisms like temporal integration of early reflections. For strong direct sound, there is a strong correlation between C80 and C50 with the DRR. However, generally, the clarity indices are less sensitive to variations of the direct sound energy at the listener's position in 6DOF, especially for sound sources with a pronounced directivity. C80 was mentioned to correlate with the auditory room size and distance perception [138]. For speech signals, C50 may be even better. From a theoretical point of view, clarity indices are also interesting in estimating the perceived distance for cases of lower direct sound. They also mirror the auditory horizon effect in auditory distance perception. Accurate distance perception and its relative change with listener motion are essential in the creation of 6DOF systems, in particular for audio-visual coherence.

Only very few studies address the estimation of JNDs for ELRs, for example, Larsen et al. [139] investigated the JND of DRR. However, such JNDs are likely to vary with the energy's temporal, spatial, and spectral distribution, which has yet to be considered in JND estimation so far. In addition, there are still debates about the criteria to estimate a suitable transition time (range) between the early and late part of the reverberation.

Generally, in the case of motion in 6DOF, there is always a relative change in the discussed parameters. This change depends on the room and the movement. This raises the question of whether the change is perceived as characteristic of a given room.

The adaptation to the acoustics of a given room also plays a role for the speech intelligibility.

Adaptation and abstraction effects

People can be experienced listeners not only due to their profession but also concerning their everyday environments like their living room at home or the office they work in. Whether a listener knows a room or has been exposed to the acoustics of a given room for at least a certain amount of time affects the perception of the reflections and rising expectations regarding the properties of the reproduced room. The results of several studies indicate that previous exposure to a room influences the perception of its acoustics. Adaption effects have repeatedly been reported in the context of speech intelligibility and echo suppression [140, 141].

The law of the first wavefront describes the effect that the localization is dominated by the direction of the direct sound, even though early reflections arrive from other directions within a short time after the direct sound. If reflections arrive later or exceed a certain energy level, they start to be perceived as separate sound events. This is described by the echo threshold. The echo threshold is not a fixed set of values but varies with various physical parameters, the type of signal, and the context. When listening to a specific early reflection pattern for some time, the echo threshold rises, while a sudden change in the pattern decreases the echo threshold significantly. Clifton explains this with expectations arising from adapting to the spatiotemporal pattern and a violation of these expectations by sudden changes [141]. She summarizes that "these expectations are most likely based on the listeners' accumulated experience in highly variable acoustic environments [...]" [141, p. 1526]. In real acoustic scenes, these changes would relate to sudden movements or a change in the room (acoustics). Keen and Freyman [142] hypothesizes that listeners form a model when they experience a sound in their surroundings. This model is quickly discarded once the acoustic environment changes. Clapp and Seeber [143] refer to this process as adaptation. When the listener can gain a deeper understanding of the room, for example, by walking around and exploring the

room acoustically, an abstraction of the room may occur. This could manifest as an improvement in speech intelligibility or other complex tasks. Experiments by Seeber et al. [144] show that previous exposure to the room increases source localization accuracy. The improvements due to one position in the room also hold for other positions and directions in the same room. Motions may help to speed up the process of understanding the scene, like estimating the size and geometry of the room and the properties of the sound sources [145].

Shinn-Cunningham [135] discussed a similar effect concerning distance perception. DRR is an essential cue for distance perception. However, in different rooms, equal DRR values correspond to different distances. Thus, the human auditory perception needs to adapt the interpretation of the acoustic cues. Motion is likely to be helpful with this as well. Klein et al. [146] showed that a previous short training in the acoustics of a room influences the evaluation of externalization. This can reduce the perceived room acoustic mismatch (room divergence) in a binaural reproduction when only virtual sound sources are audible. However, it remains open whether this training effect is still observable in augmented reality scenarios, where a real sound source is usually present throughout the AR reproduction in the given room.

In general, the adaptation process is not yet well-understood. Some experiments show high inter-individual differences in the learning process [146]. The relevant time intervals are unknown, too. In experiments with simple click intervals, adaption effects can be measured after a few hundred milliseconds (depending on the number of training clicks). For reflection suppression to increase speech intelligibility, Zahorik [147] mentions a duration of about one second while experiments with effects on externalization report adaptation over several minutes [148].

Augmented reality could be a particular case regarding the adaptation to room acoustics. In an AR application, listeners can compare the acoustics of virtual sound sources to the real ones. This side-by-side comparison makes it easy for the listener to discover differences. These ambiguities in the acoustic cues could slow down or prevent adaptation processes altogether.

2.4.3. Late reverberation

Theoretical considerations assume that at a certain time, after switching on the sound source in a room, a diffuse sound field is established. According to its definition, a diffuse sound field has a uniform sound pressure distribution and a uniform distribution of incident sound intensity. Perfectly diffuse reverberation is hardly achieved in real rooms. Romblom et al. [149, p. 1] claim that directional components in non-ideal diffuse field reverberation "may be a previously unrecognized component of spatial impression." However, starting from a certain point of time after room excitation, the listener cannot perceive direction-dependent differences. This may even hold for different positions if room modes remain negligible.

This point of time is referred to as the perceptual mixing time [150]. It can be used to simplify the synthesis of late reverberation by keeping the late part of the impulse responses constant for the different directions and possibly also for the position. Very few studies considered positional changes in determining the perceptual mixing time were conducted [151]. Pörschmann and Zebisch [84, p. 544] selected measurement positions "in the diffuse field of the sound source." While Lindau et al. [150] chose to place the speaker at twice the critical distance in the corresponding room, Meesawat and Hammershøi [151] placed them 1.5 m from the listener. This results in prominent direct sound and a high DRR compared to other positions in the room. This is not representative of 6DOF. Typically, values between 30 and 60 ms were found for the perceptual mixing time in small rooms [150, 151, 84]. Especially for 6DOF scenarios, an in-depth investigation of the perceptual mixing time that considers the occurrence of room modes is still pending.

A certain time after the start of the excitation of the room, often referred to as the physical mixing time, the reverberation can be described by a statistical time-frequency model. Such models com-

monly include parameters describing the frequency-dependent exponential decay as well as gaussian statistics of the reverberation after about 30-50 ms [152]. Examples are the spectral energy decay curves and interaural cross-correlations. Several methods have been proposed to synthesize the late reverberation tail based on the information given by an omni-directionally measured RIR. One approach is to extract the energy decay relief (EDR) and frequency-dependent decay curves. These can be scaled according to the reverberation properties of the desired room [83] or by extracting envelopes of subbands resulting from a filterbank analysis and applying them to shape a binaural noise sequence [84, 85].

Another approach is the use of feedback-delay networks (FDNs). Based on a direction-dependent target reverberation time, Alary et al. [77] create directional anisotropic reverberation with a directional FDN. Depending on the specific application, the trade-off between spatial accuracy and computational costs has to be considered. A perceptual evaluation remains open.

Regardless of whether the late reverberation is simulated or measured and adjusted, the same challenges apply to room matching. In both cases, the relevant parameters must be known to create convincing synthetic reverb or modify recorded reverb. While simulated reverberation offers more flexibility for changing specific properties, it also requires more effort to create natural-sounding reverb in the first place.

Djordjević et al. [153] conducted a modified MUSHRA listening experiment comparing the perceived naturalness of five different reverberation algorithms, including an FDN and a scattered delay network (SDN) method. SDNs, in contrast to FDNs, render the direct path component and the first-order reflections following a room model [154]. The experiment did not consider listener motion [153]. The results suggest that SDNs create a more natural-sounding reverberation than FDNs. However, the researchers considered only one specific method of FDN implementation. The underlying test method does not verify if the quality requirements for a high-quality AR reproduction are met. No results of synthetic late reverberation evaluated in an augmented reality test scenario could be found for this review. The general impression of the authors is that currently available implementations are pretty successful in adjusting the late reverberation to a given room. However, studies evaluating the suitability of AR scenarios in-depth are still pending. One challenge is the development of appropriate evaluation methods.

Reverberance can be predicted quite reliably by the parameter *perceived reverberation (pRev)* based on binaural auditory models [98, 155]. This approach can be applied directly to the audio stream without the need to extract the BRIR. The results show that reverberance correlates well with EDT. However, signal properties like the level and the spectral content have a strong influence on pRev. The results show a good alignment of the model with listening test results.

Furthermore, IACC_{Late} is associated with the perceived listener envelopment (LEV) in concert halls. It usually considers the reverberation starting from 80 ms after the direct sound. Soulodre [156], as well as Beranek [157], propose to calculate a physical measure for LEV summing a level component and a spatial component which is determined based on the IACC_{Late}. LEV has not been investigated in the context of AAR in small rooms or concerning 6DOF. A stably low interaural coherence can also indicate the diffuseness of late reverberation and the mixing time. In a coherence-based estimation of the mixing time, the moving short-time interaural cross-correlation STIACC can be helpful. However, Alary et al. [158] analyzed spatial room impulse responses (SIR) recorded with a 32-channel spherical microphone array. They determined a mixing time defined by a stable minimum of coherence and showed that after this mixing time, still directional components can be found in the reverberation of the considered concert halls. These components are audible. An investigation of the same question in small rooms would be of interest. It also remains open whether these components play a role in perceptual room matching.

2.4.4. Consideration of room modes

The eigenfrequencies (or characteristic frequencies) of a shoe box-shaped room can be calculated as follows [6]

$$f_{m,n,l} = \frac{c}{2}\sqrt{\left(\frac{m}{L_x}\right)^2 + \left(\frac{n}{L_y}\right)^2 + \left(\frac{l}{L_z}\right)^2} \tag{2.2}$$

m, n, l are integers representing the so-called quantum numbers. L_x and L_y are the distances between two opposite walls, thus length and width of the room. L_z is the height.

According to Knudsen [159, p. 36] "the qualities of all sounds, such as speech and music, are changed by the resonant properties of rooms. This change may be of a large magnitude in small rooms. Thus, certain low-frequency components which agree with natural frequencies of a room may be intensified as much as 20 to 25 dB." Knudsen also points out that the effect is robust for wavelengths in the room dimensions range.

The transition between the low frequencies, dominated by separate room modes, and the high frequencies that exhibit a dense modal overlap with Gaussian properties is smooth and continuous. Therefore, a limiting frequency can hardly be defined. Schroeder and Kuttruff [11], Schroeder [160] proposed a 3-fold modal overlap. This resulted in the definition of the well-known Schroeder frequency, which depends on the reverberation time and the room volume. It is one specific frequency value marking a region of transition. Skålevik [161] argues that the Schroeder Frequency has been "designed and tested as a low limit ensuring the validity of high frequency theory." Consequently, the value is sufficiently high, but it could be higher than necessary.

Investigations on the perception of room modes have mainly been motivated by the goal to control the modal decay for room acoustic treatment applications [162], optimal loudspeaker placement [163, 164] or the general audibility of spectral irregularities [165].

One of the recent studies concerned with the determination of perceptual thresholds of room modes was conducted by Fazenda et al. [166]. The study investigates two different perceptual thresholds as a function of modal decay. The first set of test stimuli consisted of windowed sine bursts representing the excitation of single resonances to determine absolute thresholds. The second set included music signals, considering the more complex nature of real signals in terms of temporal and tonal characteristics as they are likely to introduce different masking effects. As expected, the measured thresholds for the musical signals exhibit higher values, therefore, lower sensitivity than for the "single resonance" signals. Generally, the threshold can vary enormously with the type of signal and decreases with increasing frequency. According to the authors, there is still a lack of studies to create a coherent auditory model.

In the context of AAR, room modes have barely been considered. Many questions remain open. For example, a sufficient understanding of the listener's expectations with respect to room modes is necessary. The listener may prefer a simplified version of the sound field without considering room modes since, in many real rooms, the goal is to suppress them. Room modes are an issue for 6DOF listening scenarios, as the listener can walk through the room, and room modes can cause position-dependent fluctuations in the low frequencies. Furthermore, the perception of room modes depends not only on the listener's position but also on the source position and sound source directivity. Including these variations in the auralization would require a more sophisticated rendering of room acoustics. An understanding of the perception of modal structures in small rooms will help to optimize rendering algorithms.

One specific case is determining the mixing time to simplify late reverberation synthesis for position-dependent reproduction. Lindau et al. [150] suggested that the occurrence of audible room modes limits the extension of the mixing time concept to position-dependent reproduction. This presumably applies only to frequencies (clearly) below the Schroeder frequency.

Moreover, room modes can impose a practical issue for algorithms based on sparse positional sampling of the acoustics in a room. Either positions with strong modal effects should be avoided, or the algorithms for post-processing should be robust to their influences.

2.5. Listener translation

"Theories of spatial hearing that describe relationships between the position of the auditory event and the changes to the ear input signals during head movements are called motional or motoric theories." Blauert in "Spatial Hearing" (1997), translated by J.S. Allen, [1, p. 178]

Active head rotation facilitates the localization of sound sources in their direction [1, 167, 18, 168]. It can also help to resolve ambiguities arising from the cone-of-confusion, for example, front-back confusions [15] or above and below [169]. Even slight head movements can be beneficial in that process. A positive influence on the accuracy of auditory distance perception could not be found [170]. Hence, the question becomes interesting whether there is any (additional) benefit from translational motion, an active position change, during listening. This section discusses the role of self-translation in human auditory perception.

2.5.1. In the real world

Listening to an acoustic scene from more than one position allows one to capture more information. Each additional perspective may help to understand the current acoustic environment better. This effect is called positional disparity. The option of interactively changing the listening position and walking around also enables the listener to move closer to an object to explore it in more detail or walk around an object and listen to it from arbitrary directions.

Epstein et al. [171] propose the hypothesis that humans create a cognitive map for spatial navigation. Interactive exploration of the acoustics might be used to establish a map of the surrounding environment. Weisberg and Newcombe [172] point out that this theory has always been discussed controversially. Is a continuous change necessary for map generation, or can "teleportation" also provide this information?

Interaction with the sense of (self-)motion and the role of active translation As discussed earlier in this text, auditory perception is affected by the input of other modalities. Besides the well-known interaction with the visual sense, it is of interest whether there is a connection between the human auditory sense and the sense for (self-)motion.

When sitting on a train and watching another train starting to move, people often get the impression that their train would move. Visual information can induce the perception of motion. This effect is called vection. Vice versa, running on a treadmill is a case of self-motion without visual cues for motion. After a few seconds of such running, people walked too far when blindly walking to an object they had seen before [173]. Without running on the treadmill, people can solve this task quite accurately.

These examples show an interaction between vision and the sense of self-motion. This interaction has been subject to research for more than a century. Durgin [174] provides a review. He developed the theory that "rather than emphasizing the need for accurate absolute metrics for action [..] the precision of the relative metrics of perception and motor action are much more important". One well-known example is an observation by Harris [175]. When looking through prism glasses that cause an offset in the localization, a person can still hammer a nail after adapting to the offset. People only have to align the position of the nail and the hand holding the hammer relative to each other.

This theory could also be relevant concerning auditory perception. So far, only a few studies have looked into the interaction of the sense of self-motion and auditory perception. Auditory input can also induce vection. However, it appears to be much weaker than that caused by visual input [176, 177].

Carlile and Leung [178] reviewed a multitude of studies on the perception of auditory motion with regard to translation. The review considered the motion of sound sources observed by a static

listener and, in contrast, listeners who moved either by walking or with a motorized wheelchair. The authors point out that although in everyday life, listening during self-motion is common, most studies only considered the case of listening to an auditory motion from a stationary position.

Wallmeier and Wiegrebe [22] compared three types of rotation with a different degree of interaction, while the audio signal at the listener's ears remained the same in all conditions. The participant had to rotate until being aligned as parallel as possible to a long virtual corridor. In one case, the non-moving listener could rotate the acoustic scene around him by controlling the angular speed with a joystick. In the second test condition, the listener controlled the rotation of the chair he was sitting on with the joystick to turn around in the virtual environment. The third condition consisted of a tracked rotation of the head in addition to the rotation of the chair.

It may be concluded that actual self-rotation provides additional cues compared to being moved, e.g., by a turning chair. In both cases, the same audio signal is presented to the user. However, the proprioceptive information from self-motion provides additional cues that influence the interpretation of the heard acoustic information.

Furthermore, Genzel [179] studied the influence of a relative lateral position change between two sound sources and the listener in a distance discrimination task. In one condition, the listener moved actively. In the second condition, the listener was moved, and in the third case, the sources moved, but the listener remained in the same position. In the case of the listener's active self-motion, the smallest just-noticeable distance differences were found, and for the source motion, the results were the worst. This indicates the role of self-motion also for the case of translational movements.

Dynamic cues In addition to the acoustic information, an additional position may provide the momentary relative change of interest. In the visual domain, motion parallax effects and the time-to-contact are known dynamic cues which contribute to the perception of depth and the estimation of the egocentric distance to an object of interest [180, 181]. These dynamic cues can only be used under certain conditions, e.g., a sufficient velocity [182].

Shaw et al. [23] and Guski [24] suggested that in an approaching motion towards the sound source, the momentary change of intensity could provide additional information as a dynamic cue. This effect corresponds to the time-to-contact or time-to-collision. As an analog to the optical τ it is called *acoustic* τ. Ashmead et al. [183] compared the estimated egocentric distance of a sound source in stationary and moving conditions. The participants were blindfolded and had to report the perceived distance by walking to the estimated location. Distances between 5 m, and 19 m were considered. Significant differences could be found between listening in a stationary position, in two different stationary positions, and while walking toward the source. The authors explain the results with the effect of the acoustic τ.

Speigle and Loomis [184] conducted a similar experiment with distances of 2 m, 4 m, and 6 m. They added conditions in which the listeners did not only move towards the sound source but walked at different angles to it. In those cases, in addition to the acoustic τ, motion parallax effects may provide dynamic acoustical cues. In this experiment, the sound level of the source was not varied as in [183]. A significant difference was found for the moving condition, but the walking angle did not have a measurable effect. This experiment shows, again, that there is an influence of self-translation, but it does not show any impact of motion parallax.

It is known that trained echolocators can localize a reflecting surface, its size, shape, and material only by listening to the reflection of self-created sounds [185, 128]. Rosenblum et al. [186] observed that an active approaching motion towards the wall improved the accuracy in estimating its distance. The authors conclude that this observation is not the result of training. Instead, the acoustic τ could have contributed.

Genzel [179] claims to provide "psychophysical evidence for auditory motion parallax." In the experiment, one loudspeaker was placed behind another, and the listeners had to distinguish whether a high-pitched sound was emitted from closer or farther. At the same time, the other one produced

a low-pitched sound. In the stationary condition, the listener was placed in line with the two loudspeakers. In the moving condition, the participants were allowed to move to the side. Genzel et al. have put much effort into minimizing the known acoustical cues for distance estimation, like sound level or spectral cues. However, the potential motion parallax effect has not been separated from the disparity of subsequently listening from two stationary positions. Maybe already listening only from the lateral position might have provided better results in that task than listening from the front. Since this was not tested, it is not possible to conclude that the better results in the motion condition are due to the motion parallax effect.

Besides those few studies, there is a lack of further research confirming the effects of the acoustic τ or auditory motion parallax. Zahorik et al. [2] summarize that dynamic cues play a minor role in auditory distance estimation. For walking, one reason might be the relatively low speed. During fast position changes, potential effects of dynamic cues may be more substantial and cause measurable differences, like in [187]. This will be in line with the observation that in the visual domain, motion parallax only plays a role if a certain minimum velocity is achieved [182].

Summary The potential influences of active listener translation on the interpretation of the sound pressure at the ears can be summarized as follows:

- Additional information from positional disparity and the potential of creating a cognitive map of the environment
- Influence of the sense of self-motion
- Dynamic acoustical cues like the current change of intensity (acoustic tau) or the current change of the relative angle to a sound source or between two sources (motion parallax), maybe use of the current change of DRR

2.5.2. In dynamic binaural synthesis

For discussing the role of motion in the perception of dynamic binaural auralizations, it is important to understand the difference between the listener and the avatar representing the listener. The basic principle is illustrated in Fig. 2.6.

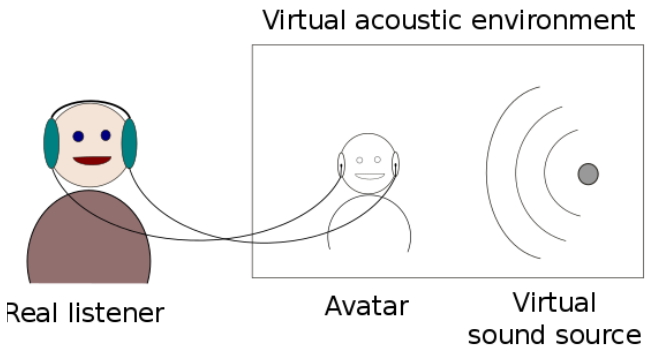


Figure 2.6.: Binaural reproduction of a virtual auditory environment

There are different kinds of interaction between the listener and the avatar. The following list provides a short overview.

- **Passive listening** - The listener only listens to the changing perspective of the moving avatar without having any control over its motion. The listener does not move or at least not in correspondence with the avatar's motion, for example as described in [188].

- **Non-authentic interaction** - The listener controls the avatar with a keyboard, joystick, or another interface. The movement of the avatar does not correspond to the movement of the real listener.
- **Authentic interaction** - The listener and the avatar move in correspondence, realized with motion capture. Another degree of authentic interaction is achieved if the listener and the avatar create the same sounds, e.g., if the listener lets the avatar speak with his own voice.

There are also various **Subtypes**, for example, controlling the position of an avatar with keyboard and controlling orientation with tracked own head rotations. An extended discussion of the types of interaction is provided in [145].

For systems with non-authentic or even no interaction between listener and avatar, the question for the occurrence of auditorily induced vection may be interesting. Boerum et al. [189] studied this question by asking seated listeners whether they had the impression of self-motion when they were listening to a continuous change of the observer position.

As discussed in the previous subsection, the auditory perception of a real acoustic environment depends on the degree of active motion in the exploration. Consequently, a system for creating virtual acoustic environments or virtual sound field elements should provide the option of an interactive exploration in 6DOF.

Generally, spatial auditory illusions will be perceived as more plausible, more immersive, more vivid, and more interesting if they endure listener motion. Therefore, the consideration of listener motion with authentic interaction between the listener and the avatar representing it is of interest. The investigations within this thesis consider only authentic interaction.

If a system is intended for the realization of authentic interaction with a system, a careful choice of the system parameters is essential. A physically perfect authentic interaction would be achieved if the avatar would be a perfect replicate of the listener and do the same movements at precisely the same point of time as the listener and if the acoustical information at the eardrums could be recreated in every detail. Real systems will always exhibit deviations from this optimum due to unavoidable system delays and approximations of the individual head and body shape. For this reason, it is necessary to understand the just-noticeable-extend of these deviations and at which point will affect the quality of the auditory illusions.

Integrating the option of interactive head rotation in binaural systems was already shown to improve externalization [19], reduce front/back confusion, and improves localization of source direction. Although no influence of head rotation on auditory distance perception was found in the real world, Kearney et al. [190] investigated the same question in a dynamic binaural auralization. Again, no significant impact was observed.

So far, there are only a few studies on the perception of the perspective of virtual listeners that move through their virtual auditory environment. It is a complex perceptual process. The interaction with other modalities is not well-understood. The type of interaction with the avatar will influence how the user perceives the acoustical scene. One crucial question is the required resolution of the sound field parameters subject to change during listener translation. This is, for example, essential for the determination of required update rates.

Studies on the perception of dynamic binaural synthesis for head rotation usually focus on certain listening positions, like the sweet spot of a virtual loudspeaker system or a scenario in which a virtual person is located nearby but not too close and speaks directly towards the listening position. Reproduction systems that allow for an interactive exploration by walking around invite the user to move to less common listening positions. These could be in front of a source, but also behind it, far away from it, in the corner of a room, close to a wall, or very close to a source. In order to gain solid background knowledge about the auditory perception in a 6DOF exploration of a room, these cases have to be studied.

2.6. Different approaches to realize position dynamic binaural auralizations

This section is based on the introduction of the journal article (preprint) "Effect of impaired early reflection patterns on plausibility and similarity of position-dynamic binaural AR audio" [314] that I submitted for publication.

For position-dynamic binaural audio, the BRIRs must be adjusted without a noticeable delay to the source's or listener's current position and orientation relative to the sound source and the room. Updates are required in a sufficient spatial resolution or even continuously. The generation of a satisfactory BRIR data set, either by measurement or simulation, can be tedious. Researchers are investigating different approaches of parameterization, interpolation, and extrapolation to create efficient binaural audio systems for AR/VR/MR, for example by Arend et al. [85], Werner et al. [191], Tylka and Choueiri [192], Plinge et al. [193], Müller and Zotter [69], Coleman et al. [82], McCormack et al. [70]. High physical accuracy usually demands high computational capacity and much memory space. However, listeners exhibit limited sensitivity to the physical details of a sound field. Therefore, psychoacoustic optimization is of interest and is currently subject to research.

A well-known example is perceptual mixing time. It marks the time after exciting the room, from which the reverberant sound field is perceived as diffuse and, thus, independent of position and direction. This is interesting for dynamic binaural reproduction, as the late reverberation may be modeled with a static transfer function. This can increase the efficiency substantially, in particular for rooms with a long reverberation time. Lindau et al. [150] investigated the perceptual mixing time t_{mp} for nine rooms with varying volume and average absorption in a psychoacoustic experiment considering head rotation in the range of $\pm 80°$ azimuth, but no position changes. The adaptive 3-AFC method was applied to test for audible differences. For the three small rooms in the sample, the mean perceptual mixing time was earlier than 50 ms after the direct sound. The values increased with the size of the room. Linear models were derived from the empirical data to predict the perceptual mixing time based on room volume V and reflecting surface S. Additionally, signal-based models for estimation from a measured room impulse response were developed. For each of the nine rooms, only one receiver position was taken into account. Moreover, the audio content plays a role in the discriminability of room reflections [194]. Lindau et al. [150, p. 893] used a 'critical drum set sample' with a duration of 2.5 s, which was reproduced only once for each condition. An omnidirectional sound source was used for the acoustic measurement.

Stade [195] evaluated the similarity of BRIRs with constant reverberation tail after different transition times to reference BRIRs which were measured in two rooms ($V \approx 1200\,m^3$, $RT \approx 0.8\,s$ and $V \approx 6000\,m^3$, $RT \approx 1.5\,s$) at single selected positions for head rotation. The test was conducted once with the full BRIRs and once, the early part before the transition time was removed. Without the early part, audible differences to the reference were substantially more distinct and still observable even for very late transition times, where the complete BRIRs could not be distinguished. This suggests a masking effect of the early part concerning deviations in the late reverberation.

Meesawat and Hammershøi [151] studied the audibility of differences when the late reverberation was taken from a different position in the room. For the chosen room ($V = 185\,m^3$, $RT = 0.64\,s$), transition times of 40-60 ms seemed appropriate. However, only static binaural synthesis was taken into account, and all considered listening positions were located 1.5 m in front of the loudspeaker. This leads to strong (masking) direct sound arriving at the listener. Lindau et al. [150] noted that replacing the late reverberation tail with one measured at a different position in the room caused more significant differences in the low frequencies. This leads to audible differences also in case of late transition times. They argued that the concept of perceptual mixing could not be extended to position changes because room modes are likely to cause position-dependent audible effects. In small rooms, room modes are generally of higher practical and perceptual relevance [6]. A perceptual evaluation of the perceptual mixing time in position-dynamic binaural reproduction is still pending.

Besides physical simplifications after the perceptual mixing time, a perceptual optimization of the early reflections (ER) is of interest. The accurate rendering of ER requires very accurate information about the geometrical arrangement of the environment, the acoustic properties of each surface, the type and directivity of the sound source, as well as the positions and orientations of sources and the listener in the room. This information is often not available, for example, in AR systems used in unknown environments. Furthermore, if it is available, considering all of this information in dynamic binaural rendering can be very costly.

For a single reflection to cause any audible effects, it must exceed the *audibility threshold*. Thresholds of -10 dB with regard to the direct sound if it comes from a similar direction, to below -20 dB for other directions were observed by Olive and Toole [194], Buchholz et al. [102], Zhong et al. [103]. Rakerd et al. [196] reported an average threshold slope of -0.20 dB/s with increasing delay. Besides the angle of incidence, relative delay, and level with respect to the direct sound, the threshold depends on the type of signal with its frequency range. Moreover, the perception of a single reflection is influenced by further occurring reflections and the late reverberation, as shown by Bech [163, 119, 120].

In literature about the precedence effect, it seems widely accepted that summing localization causing an apparent directional shift of the source image, occurs at very low delays of up to 1-2 ms, with a more complex signal like speech or music up to 5-10 ms. However, Barron [115] documented image shift effects for larger delays, particularly if the reflection was stronger than the direct sound. Olive and Toole [194] observed an *image shift threshold* of about 7 dB above the audibility threshold for reflections delayed by 30 ms, arriving from 65°, considering also image spread effects.

Hacıhabiboğlu and Murtagh [132] showed that the number of auralized reflections could be reduced considerably without affecting localization and other spatial qualities of the reproduction. The selection of the ERs is based on the time of arrival and the angle of incidence. Similarly, Brinkmann et al. [80] realized a parametric encoder based on a determined minimal set perceptually salient ER. The experiment considered hybrid simulations of omnidirectional sound sources in empty shoe-box rooms of different sizes and with different reverberation times. The threshold -10 dB relative to the direct sound and with a decay of 1 dB/ms for the frontal direction and a further direction-dependent decrease for lateral reflections. The receiver was simulated at twice the critical distance from the source. For these test cases, encoding the six first-order reflections was sufficient to create indiscernible auralizations. These findings demand an evaluation in more realistic scenarios.

Zotter and Frank [121] demonstrated that the first-order reflections of a simulated rotating sound source with third-order directivity could have a substantial influence on the localization of the apparent source image. They also showed that adding second-order reflections did not have a considerable influence on source localization in their test cases but smoothed the transition of the apparent source location during source rotation. According to Steffens et al. [122], the directivity of a sound source should be considered in the first-order reflections. The source orientation remains distinguishable - at least for angle changes of 90° if reflections of second order and high are simplified fixed average directivity filters. Further minimization of coloration effects can be achieved by applying the fixed average filter starting from the third order.

Arend et al. [85] propose a method to create a binaural 6DOF auralizations based on the parametric extrapolation from an omni-directionally measured single-channel room impulse response (mono-RIR), extending the ideas proposed by Pörschmann et al. [63]. The method can be realized with different approaches to reconstructing binaural ER patterns from a RIR. An SDM-based DOA estimation allows an approximate reconstruction of the original DOAs as well as times-of-arrival (TOA) and the energetic relations of the ER. An extrapolation for different listening positions could be achieved by manipulation of the ER pattern in accordance with an assumed shoe-box room model or by simply assigning pseudo-randomized directions-of-arrival (DOAs) to the reflections. The perceptual consequences of unnatural spatiotemporal ER patterns have not been understood in depth yet. A thorough psychoacoustic examination is of interest.

For an Ambisonics-based representation, the sound field is encoded in the spherical harmonics

domain [197, 198]. In a *standard ambisonics* realization, a truncation of the order will lead to an approximation or simplification of the sound field's physical details, which increases with frequency. A lower maximum order decreases the computational load but also the spatial resolution of the representation. Engel et al. [199] discusses *hybrid ambisonics* or the *reverberant virtual loudspeaker (RVL)*, two further approaches of ambisonics-based simplifications that can seriously affect the natural progress of early reflections. For the realization of ambisonics-based 6DOF-auralizations, different approaches of interpolating or extrapolation from a combination of several ambisonics recordings with distributed microphone arrays have been proposed, for example, by Tylka and Choueiri [192], Patricio et al. [200], Plinge et al. [193]. Generally, maintaining the natural localization and coloration as well as a smooth transition of both between different listening positions is challenging, but according to Müller and Zotter [69], considering the nine most prominent reflections was beneficial in this context.

Different realizations can vary substantially in the achieved quality. Therefore, strategies to evaluate and measure the quality of such systems will play an essential role in this thesis and are therefore presented and described.

2.7. Plausibility of auditory illusions for VR and AR

This section is based on the introduction of "The availability of a real hidden reference affects the plausibility of position-dynamic auditory AR" [275] that I published together with Anna Maria Zerlik in "Frontiers in Virtual Reality" in 2021.

With the goal to realize auditory illusions with low-cost devices, for example, as described by Heller et al. [201], there is the desire to identify the potential for optimization without affecting the quality of the resulting spatial auditory illusions. This process demands appropriate methods to evaluate the achieved quality. One essential question is how the created virtual acoustic object perceptually compares to the corresponding real version if there is a real version. In this context, *authenticity* and *plausibility* have become essential constructs.

According to Blauert [1], *authenticity* describes the agreement of the perceived acoustical scene with an external reference. Thus, a virtual acoustic object created with binaural reproduction is considered authentic if it cannot be distinguished from the corresponding real version in a direct comparison.

Slater [202, 203] proposed the *plausibility illusion* as one of the critical components in the perception of multi-modal VR realizations. He linked this term to the overall credibility of the scenario compared to a user's expectations. While sticking to this basic understanding, Kuhn-Rahloff [204] adopted the construct for the evaluation of auditory contents. According to this proposal, *plausibility* describes the agreement of the perceived acoustic scene with the listener's internal reference. This internal reference is the expectation that results from a person's individual listening experience.
Latoschik and Wienrich [205, p. 5] argue that in AR, "the central idea is to augment a physical space with additional computer-generated entities and not to artificially simulate a virtual space." Rather than assuming an illusion of plausibility like Slater [202] and Skarbez et al. [206], they define plausibility as "a state or condition during an XR experience that subjectively results from the evaluation of any information processed by the sensory, perceptual, and cognitive layers." In addition, Latoschik and Wienrich [205] propose a novel model describing XR experiences and effects wherein coherence and plausibility constitute central essential components. This model is still based on the idea that perceptual cues, sensory cues, and higher-order (cognitive) cues have to be in line with the experience and expectation of the user to achieve coherence and plausibility.
According to all these definitions, a virtual acoustic object is plausible if it fulfills the listener's expectations. Slater [202] and Skarbez et al. [206] state that a virtual element can be plausible

even if the user knows it is not real. However, suppose a virtual replicate of a real sound object is in satisfactory agreement with the individual expectations of a listener. In that case, this listener will not be able to surely tell that the acoustic object is virtual and will accept it as real. At this point, the highest degree of plausibility is achieved. If the internal reference is of limited accuracy, the listener may accept an inaccurate virtual replicate as real. In contrast, listeners with a wrong internal reference may not even accept the real version as real. One of the challenges in evaluating plausibility is the limited reliability and stability of a listener's internal reference.

Several studies assessed the authenticity of spatial auditory illusions created with static binaural synthesis without the option of interactive listener motion, for example, Moore et al. [207], Maseiro [208], Oberem et al. [209]. Brinkmann et al. [29] presented the first study investigating the authenticity of virtual sound sources in different real rooms created with dynamic binaural synthesis considering interactive head rotation. For the realization, a simulated equivalent of a real scene is created for dynamic reproduction based on individual BRIR measurements. For these measurements, extra-aural headphones [210] were placed over the ears of the listener to consider their influence in listening to the real scene. An experiment with an individual Two-alternative-forced-choice (2AFC) test paradigm was conducted to test for minor noticeable differences. With the given realization, an authentic, dynamic binaural reproduction for interactive head rotation was achieved for the speech signal but not for the noise signal.

An authentic implementation demands high technical precision and effort. In AAR, usually, a direct comparison to the real version is not possible. Thus, for many applications, the concept of plausibility is more interesting. Lindau and Weinzierl [211] proposed a method based on the signal detection theory to evaluate the plausibility of a dynamic binaural synthesis system. Again, a real sound field and its binaural simulation are considered. In the experiment, randomly, either the real scene or the binaural auralization was provided to the subjects. They had to decide in a Yes/No paradigm which version they were listening to. The basic idea of using a Yes/No paradigm in a mixture of real and virtual sound sources was not new at that point. This approach was employed, for example, by Hartmann and Wittenberg [212] to evaluate *externalization* and *convincingness*, by Langendijk and Bronkhorst [213] to investigate the *fidelity* of virtual sound sources and in an earlier study [214]. However, Lindau and Weinzierl [211] took this approach to a new level of depth and linked it to plausibility as proposed by Kuhn-Rahloff [204].

Including a real sound source as a test case in an experiment requires considering how the presence of the headphones affects the perception of the real sound source. This effect is added to the virtual version to avoid this occlusion or shadowing effect, causing audible cues only for the real scene. A new set of BRIRs has to be measured with the desired pair of headphones placed on the listener's or the dummy's head. Investigating a 6DOF system causes considerable effort because each position of interest has to be measured separately. Moreover, a slightly distorted perception of the real sound source caused by the occlusion can lead to additional confusion. On the one hand, listeners could increasingly mistake the real sound source for the virtual version. On the other hand, this approach can only investigate the quality of a spatial auditory illusion of a slightly distorted reality. This is a common challenge in realizing AAR systems, which provide virtual content alongside the real acoustic environment. Is it a suitable approach to encourage the creation of virtual content containing the same effect?

The method suggested by Lindau and Weinzierl [211] is valid and interesting for evaluating the reproduction system itself. However, not only reproduction systems need to be tested for plausibility, but also fictional scenes or other content for which there is no real counterpart. If the scene contains a cartoon hero or a little ghost flying around or if a product is designed virtually and realized later on, how can we evaluate the plausibility in such cases? These questions are also interesting for Virtual Reality, where the listener can be transferred to a fantasy room like in the studies by Enge et al. [215] or Remaggi et al. [316].
In the field of VR, scientists have started to distinguish between *internal* and *external plausibility*. Hofer et al. [216, p. 2] provide a nice summary of that discussion. In this understanding, internal

plausibility "refers to the extent to which the environment is consistent within itself or with respect to the expectations raised by its genre." An example of violated internal plausibility, as defined by Hofer et al. [216], would be to have a vegetarian that eats meat in the scene because new information - the character eats meat - contradicts already presented information - the character is a vegetarian. External plausibility in this context "refers to how consistent the virtual environment is to users' real-world knowledge" [p. 2]. This definition addresses whether the presented scenario could occur in the real world, but it is not necessarily indistinguishable from reality. These interpretations and classifications of plausibility refer to the credibility and consistency of the content rather than the rendering quality, which we consider in our discussion of plausibility. Our study only considers scenes that can occur in the real world, which is external plausibility as Hofer et al. [216] describe it. Still, it is essential to note that methods to evaluate plausibility based on a comparison with a real counterpart have the limitation of not being helpful for fictional content.

Binaural Synthesis	**Plausibility I** pure internal reference	**Plausibility II** 'tuned' internal reference	**Authenticity** external reference
Static Reproduction	(✓)	Hartmann & Wittenberg 'Convincingness' (1996) Oberem et al. (2016) part B	Moore et al. (2010), Maseiro (2012), Oberem et al. (2016) part A
Head rotation	(✓)	Lindau et al. (2007), Lindau & Weinzierl (2012), Pike et al. (2014)	Brinkmann et al. (2017)
Rotation & Translation	?	?	?

Table 2.1.: Summary of previous studies investigating plausibility and authenticity of binaural synthesis. Plausibility is split up into the two proposed categories of measuring the agreement with the pure internal reference or a tuned internal reference as a result of the indirect comparison with the real counterpart of the scene. This overview is not exhaustive but provides examples for each of the cases.

So far, it has not been investigated whether including a real sound field in the test paradigm would influence the result. If that is the case, it may be valuable to distinguish different kinds of plausibility, for example, indicating the agreement with the **pure internal reference** or the **tuned internal reference** as a result of listening to the real version of the scenario. Table 2.1 summarizes a selection of previous studies on the authenticity and the two proposed categories of the plausibility of auditory illusions created with binaural technology. In addition, we ordered the studies by the considered degree of interactivity. In a static reproduction, no interactive motion is possible. Several studies already took interactive head rotation into account. The option to interactively walk to another position relative to the virtual sound source is still a pretty new challenge concerning the evaluation of plausibility.

A potential tuning of the internal reference may not only occur in an indirect comparison with the real counterpart. Especially for AAR, the actual environment and its components will likely influence the internal reference. Since the scenario allows for a direct comparison, maybe the term mixed reference is more appropriate in this case. Wirler et al. [217, p. 1] proposed the concept of *transfer-plausibility* as the *"ability of a virtualized source to stand alongside multiple real sound sources"* and studied the plausibility of virtual sound sources in real environments under varying scene complexity in terms of the number of concurrent loudspeaker signals. The setup realized dynamic binaural synthesis with 6DOF, but the participants were seated during the experiment. Their results suggest that an increased scene complexity decreases the number of correctly identified virtual sound sources, even with a rendering of lower quality. The concept of *Co-immersion* proposed

by Stecker et al. [218] addresses this topic similarly.
It is likely that not only the number of sources or the scene complexity have an influence on the internal reference but also the type and the relative positions of the available real sound sources. If, for example, a virtual loudspeaker is created next to a real loudspeaker, achieving a quality of the illusion that listeners cannot identify as virtual may be more challenging than if the sound of a person riding a bicycle is added to an acoustic environment with a distant street full of cars.

3. Implementing a setup for position dynamic binaural auralization

The first step of this thesis is the realization of a setup for position dynamic binaural auralization of sound sources in a room. The listener can explore the acoustic scene by interactive own position changes.

This chapter documents and explains the technical properties of the implemented system and the devices used within the experiments. First, the development of the Python tool pyBinSim for dynamic binaural rendering is described, followed by the procedures to create BRIR data sets for listener translation based on acoustic measurements and room acoustic simulation. Afterward, the system implemented for the dynamic binaural audio during the experiments is presented, and finally, a first approach informal evaluation of plausibility is documented.

3.1. PyBinSim - a flexible python tool for dynamic binaural synthesis applications

This section is an extended version of the engineering brief e-brief346 "Flexible python tool for dynamic binaural synthesis applications" [317] that I published together with Florian Klein, Niklas Knoop and Thomas Köllmer at the 142th AES Convention in Berlin, Germany, in May 2017.

This section describes the development of a light-weight tool for dynamic binaural synthesis developed for the investigations in this thesis, namely *pyBinSim*[1]. First, an overview of tools existing at the beginning of 2016 is provided, followed by a summary of state-of-the-art algorithms. After that, the design and implementation of pyBinSim, as well as application examples, are explained. The tool has evolved, and new features have also been implemented in 2022. More information is provided in sec. 5.2.

Available tools and motivation to develop a new tool in Python The SoundScape Renderer (SSR) is a well-known open-source renderer for spatial audio [219] [220]. First, it was developed as a tool for wave field synthesis (WFS) auralizations, but it also contains two modules for dynamic binaural synthesis. At that time, both modules did not allow for position tracking. Additionally, the SSR could be used only on Linux and OS X for a long time. The community was working on a comfortable solution for Windows-based systems. On top of the SSR, Scale [221] was designed to carry out listening tests considering head rotation. It allowed comparative switching between different filter sets.
Within the BiLi-Project, an experimental mixing tool called *bipan* [222] was developed. It was based on a custom version of the *Spat* binaural engine from IRCAM. As a follow-up, with myBino [223], a Vst-Plugin for dynamic binaural monitoring has been published.
Another convenient way to create acoustic scenes for VR was gaming engines like Unity which often used *FMOD engine* [224] as the primary sound-effects system to spatialize audio. Using custom BRIRs was usually not possible with the game audio engines. Furthermore, the exact sound processing within these engines was not documented publicly in detail necessary for scientific investigations.

[1] pypi.org/project/pybinsim or github.com/pyBinSim

Another tool available was the *ambiX suite* [225], which allowed a dynamic auralization of spherical sound field (Ambisonics) data over loudspeakers or headphones.

At that time, more and more institutions were developing tools for a convenient dynamic auralization of spatial audio content. The main motivation to develop pyBinSim was the need for a simple BRIR convolution tool to investigate listener movement in a virtual acoustic space based on measurements in real rooms. The tool should be independent of the devices used for listener tracking by employing an interface based on Open Sound Control (OSC). Furthermore, the desired tool can be easily used on Windows-based PCs. It was also of interest for the education of students. Python was chosen for the tool, because it is easy to learn and very flexible.

Algorithms for real-time binaural synthesis Binaural synthesis is based on applying direction- and position-dependent FIR filters to a dry mono audio signal. The filter is chosen according to the motion of sources and listeners. For a continuous reproduction that allows real-time interaction without a perceivable delay, the audio signal and the filters must be divided into small chunks or blocks.

Frank Wefers [62] provided a detailed overview of the different approaches to implement partitioned convolution and discusses their individual efficiency as well as other practical issues.
The presented implementation is based on a uniformly partitioned convolution using the overlap-save approach (UPOLS). With non-uniform partitions, variable block size can help decrease the computational effort significantly [62]. This might be included in pyBinSim in the future.

Based on psychoacoustic investigations, it has been reported that a latency below 58 ms does not cause audible effects [59]. One of the main factors of the system delay is the update rate of the tracking device. Additionally, the block size has an important influence on the overall system latency because the changes in head position or orientation will be taken into account in the upcoming block but not in the block reproduced at that moment. The maximum delay caused by the block size can be calculated according to equation (3.1).

$$\Delta t_{max} = n_{block}/f_s \tag{3.1}$$

For example, a block size of 256 samples and a sampling frequency of 44.1 kHz result in a maximum delay of $\Delta t = 5.8ms$. Additionally, further sources of delay have to be considered.

3.1.1. Design of pyBinSim

Basic principle Depending on the number of input channels (wave-file input or live audio input), the corresponding number of virtual sound sources is created. The filter for each sound source can be selected and activated via OSC messages. The messages contain the number index of the source for which the filter should be switched and an identifier string to address the correct filter. The correspondence between the parameter value and filter is determined by a filter list which can be adjusted individually for the specific use case.

The simple filter assignment provides high flexibility when using this tool for your project. However, if only horizontal head rotation should be taken into account, it is advisable to use the azimuth angle as a parameter. If the goal is to switch virtual rooms, one parameter could be used to switch between different filter sets. For applications considering position-tracking and head-tracking, the six parameters could address the 6DOF (six degrees of freedom): x, y, z, yaw, pitch, and roll.

A headphone equalization filter can be applied in real-time by activating that option. The filter has to be provided by the user. As another optional setting, time-domain-cross-fading between two blocks when switching filters can be activated or deactivated.

Integration of tracking data The integration of tracking data is realized with the OSC interface described before. However, when the listener moves or turns the head, the filters of all sources have to be updated. Since pyBinSim does not provide signal processing except the real-time convolution and the optional headphone equalization, it is up to the user to provide the correct filters.

Examples of simple integration of the tracking data, for example, from Oculus Rift and HTC Vive, are provided within the pyBinSim package.

System requirements PyBinSim runs on different platforms without having too many dependencies. The signal processing is based on pyFFTW [226] (which requires numpy). Audio output is reproduced via pyAudio [227]. Tab. 3.1 lists the core requirements of pyBinSim and its officially supported platforms.

Table 3.1.: Main dependencies of pyBinSim and availability of official packages (as of May 2017).

Library	Python Support	OS Support
pyFFTW	Python 3.4 - 3.5	Win, Linux
pyAudio	Python 3.4 - 3.6	Win, Linux, OS X
numpy	Python 3.4 - 3.6	Win, Linux, OS X
pySoundfile	Python 3.4 - 3.6	Win, Linux, OS X
python-OSC	Python 3.4 - 3.6	Win, Linux, OS X

PyBinSim itself is developed using the Anaconda environment on Windows and Linux, using Python 3.5. Other environments are expected to work as well, for example, pyBinSim on OS X using Python 3.6, Python 3.4, or even newer versions of Python.

The different trackers are integrated using a separate application which transforms the tracking data into OSC messages, that are sent to pyBinSim. Usually they require further packages or APIs:

- Razor-AHRS [228] - read from COM-Port e.g. via *pyserial*
- Oculus Rift - read from API available with *ovr*
- HTC Vive - read from API available with *openvr*
- UDP communication for trackers connected via ethernet, for example, via *socket* from the standard library

3.1.2. Evaluation: Processing performance and open issues

To evaluate the performance of the current implementation, pyBinSim was tested on two computer systems:

1. Intel Core i7-6700K (4*4GHz); Windows 10
2. Intel Core 2 Duo E8400 (2 x 3GHz); Windows 7

The performance test was conducted in a best-case scenario, meaning no filter switching was considered, and headphone equalization was turned off. Fig. 3.1 shows the dependencies between the number of sound sources and the maximum filter length per source for both computer systems. The maximum filter length is expressed as a power of 2. The graphs show how the filter length has to be reduced with increasing virtual sources for each block size to avoid dropouts. The performance advantage of the Intel i7 is due to the better performance of the single core and not due to the increased number of cores.

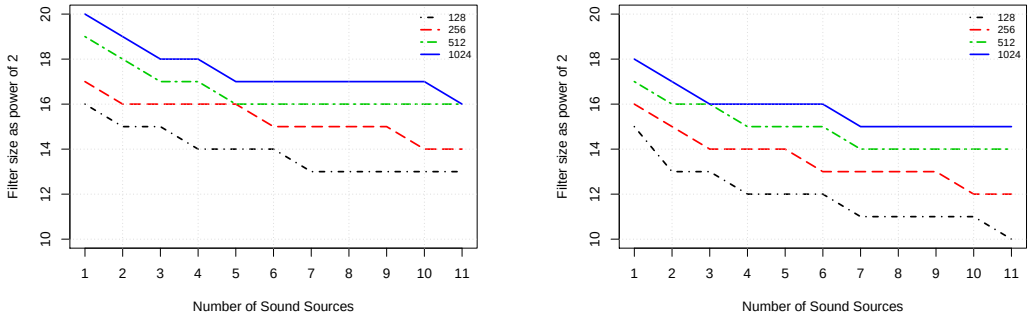


Figure 3.1.: *Left:* Maximum filter length depending on the number of sound sources for four different block sizes on an Intel Core i7-6700K (4*4GHz); Windows 10 - *Right:* Maximum filter length depending on the number of sound sources for four different block sizes on an Intel Core 2 Duo E8400 (2 x 3GHz); Windows 7

This means there is still lots of potentials to improve the efficiency. The convolver class has been implemented using the multi-threaded package in order to achieve an efficient solution. However, *cPython* does not allow the distribution of threads over the different CPU cores because of the global interpreter lock (GIL). This limits the processing power. Using multiprocessing or a different Python implementation like *grumpy* or *jython* could provide a solution.

It is recommended to use a 64 bit Python version on a 64-bit CPU platform to avoid memory issues.

Currently, filters need to be provided as separate Wave files registered in a filter list text file. However, it is interesting to allow better handling of complex filter sets in future releases, for example, by reading from SOFA-Files [229].

Application examples PyBinSim can be used as is with basic tracking solutions which are provided in the repository. Furthermore, it allows for designing own applications, even in other programming languages like C, Pure Data, or Max MSP. Some examples are given below.

- **BRIR-based auralizations** - PyBinSim allows any dynamic auralization based on an available filter set for discrete positions and head orientations. Changes in elevation can be considered with the current version, if corresponding BRIR filters are available, and if the tracking unit can provide the corresponding values.

  A virtual loudspeaker setup can be realized within the processing capabilities of the used computer (considering a single-core CPU performance). Fig. 3.1 gives an idea of the limitations at the time of its release. PyBinSim might also be of interest for the dynamic auralization of spherical microphone array recordings. This is possible using a spherical setup of virtual loudspeakers or extracting directional FIR filters, like BRIRs.

- **Include position tracking** - For translational movements of a sound source or a listener, various approaches to audio rendering exist. A BRIR-based synthesis can be realized easily. Advanced interpolation methods are currently not provided but can be integrated individually.

  Furthermore, designing an application following the object-based approach is possible. Applying distance-dependent loudness and delay, as well as the Doppler effect, might be interesting. So far, such an application has not been implemented, but it could be built on top of pyBinSim.

- **Switching between acoustic scenes with head rotation** - This basic application reads the tracking data from a particular tracking module and sends the azimuth data and the filter set

index via OSC to the rendering unit. For example, the user can switch between different filter sets for direct comparison by using specific keys on the keyboard. On top of this basic setup, listening test software, for example, for MUSHRA or ABX tests, could be created.

3.1.3. Summary

With *pyBinSim* a Python tool for creating various applications requiring real-time binaural synthesis has been presented. On the one hand, it is a ready-to-use tool for dynamic binaural reproduction. The desired filters can be loaded and assigned easily. Example code for some standard tracking solutions is provided. On the other hand, various more complex applications could be designed on top of the provided Python classes. Since the convolution can be controlled easily via OSC messages, pyBinSim could be combined with room simulation tools or other applications written in other programming languages.

Due to its simple structure, its flexibility, and lightweight design, pyBinSim offers great value to research and education in the field of dynamic binaural reproduction and interactive virtual environments.

The source code is available under the MIT license to citeMITLicenseOpenSourceOrg on Github:

```
github.com/pyBinSim/pyBinSim
```

The following two sections will describe the creation of BRIR data sets for position dynamic binaural synthesis.

3.2. Creating BRIR data sets for listener translation based on measurements

For the measurement of the first BRIR data set for listener translation, two loudspeakers Genelec 1030A [230] shown in Fig. 3.2 were set up in the listening laboratory of the university in Ilmenau. The room has a size of 8.4 m×7.6 m×2.8 m, a volume of V = 179 m^3 and a reverberation time

Figure 3.2.: Loudspeaker Genelec 1030A (picture taken from [231]).

T_{30} = 0.23 s. It meets the requirements suggested by recommendation ITU-R BS.1116-3 [232]. These demand very low background noise and a short reverberation time. Furthermore, the recommendation requests that early reflections reaching the listening area „up to 15 ms after the direct sound, should be attenuated by at least 10 dB relative to the direct sound in the range 1-8 kHz“ [232, p. 15].

A head-and-torso-simulator (HATS) of the type *Knowles Electronic Manikin for Acoustic Research* KEMAR 45BA from G.R.A.S. Sound & Vibration was used as the receiver. It is shown in Fig. 3.3.

Brüel & Kjær 4134 microphones capture the sound field at the entrance of the KEMAR's ears. Throughout all measurements, the KEMAR was equipped with small ears (KB0060/61).
The KEMAR has been designed with median human adult dimensions. The ear simulation matches the acoustic response with an auricle, an ear canal, and an eardrum that equal the median ear in dimensions, acoustic impedance, and modes. The KEMAR was standardized by ANSI in 1976, followed by the IEC and later ISO.

Figure 3.3.: The KEMAR 45BA head-and-torso-simulator from G.R.A.S. Sound & Vibration

The Kemar was subsequently placed at nine potential listening positions along a line with a length of 2 m leading towards the frontal loudspeaker. The measurement positions were arranged in intervals of 25 cm, covering the distances from 1.25 to 3.25 m from the center of the loudspeaker (approx. 1.1 m to 3.1 m from the membrane). In this first data set, this distance of 25 cm was chosen because it seemed reasonable. An evaluation of a suitable positional resolution was conducted later on and is described in Sec. 4.3. The second loudspeaker was placed at the side of the line. The HATS' ears were located at a height of 1.59 m above the floor, the same height as the acoustic center of the Genelec 1030A loudspeakers.

An electronic turntable Outline ET250-3D [2] ensured an accurate rotation of the HATS in 4° steps. Consequently, 90 BRIRs were measured at each of the nine positions with two additional BRIRs for azimuth angles of 90° and 270°. Fig. 3.4 illustrates the setup in the room. The swept sine method [52] with exponential sine sweeps ranging from 50 Hz to 20 kHz over a duration of 3 s was used. The corresponding measurement software was implemented in Matlab. Sound in- and output were connected with the PC via an RME Fireface 400.

This BRIR set is part of a more extensive data set, presented in [318] and is available online [319]. In addition, a detailed physical analysis is provided in the appendix.

3.3. Creating BRIR data sets for listener translation based on acoustic simulations

In 2016, very few room simulation tools were available as open-source projects or freeware. Furthermore, using a custom HRTF data set with a high spatial resolution was of interest. This was not available at that time in commercial room simulation tools.

MCRoomSim [233] is a toolbox for room simulation based on ray-tracing. It allows for modeling shoebox-shaped rooms without furniture, but it enables the user to customize the directivity of sound sources and receivers. Thus, it is possible to simulate virtual dummy head recordings with

[2] https://outline.it/outline-products/legacy/measurement/et-250-3d/

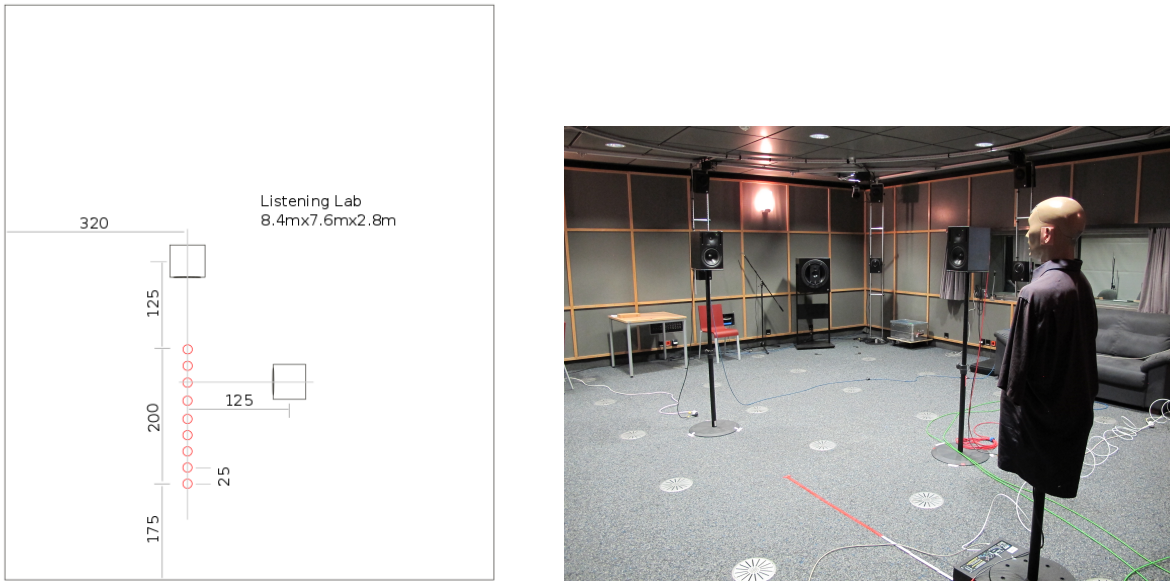


Figure 3.4.: Arrangement of the measurement positions along a 2 m line leading towards/past the two loudspeakers in the listening laboratory. BRIRs were measured at positions in intervals of 25 cm a nd for an azimuth resolution of 4°.

a desired HRTF data set. In this modeling process, HRTFs and source directivity are applied as direction-dependent FIR filters according to the nearest-neighbor principle in the rendering process.

For creating the BRIRs for this study, a sound source with the directivity of the loudspeaker *Tannoy V6*, created from a dataset provided within the MCRoomSim toolbox, was placed in the virtual room. Additionally, two receivers were positioned at the center of the virtual head position. HRTFs measured with a Neumann KU100 dummy head [234] were used to model the directivity of the two human ears.

To generate a BRIR data set for listener translation, a room similar to the listening laboratory with the same size was created. Two sets of absorption coefficients were chosen to achieve a reverberation time similar to that of the listening laboratory and create a more reverberant room of the same size. The BRIRs were generated for positions equivalent to the measured ones in the lab, with an azimuth resolution of 4° and distances of 25 cm along the 2 m line for listener translation.

No individualization is applied to the generated BRIR data. Placing the two receivers representing the ears in the same position brings along some minor deviations from real physics because, in reality, the two ears "sample" the sound field at two different positions. However, at a certain distance to the sound source and the walls, the deviations are assumed to be negligible with regard to perception. Furthermore, with MCRoomSim, rigid bodies like the human head cannot be taken into account. Consequently, the placement of both ear receivers in the actual interaural distance would not result in a more accurate simulation.

3.4. Position dynamic binaural auralization of room acoustics - The reproduction setup

3.4.1. Motion tracking and Head mounted displays

In all experiments for this thesis, motion tracking was realized with the HTC Vive tracking unit based on two Lighthouse base stations. In the early studies, the participants had to wear the HTC Vive head-mounted display to capture the orientation and position of the head. Later, the HTC Vive tracker was also available that could be attached to the headphones and enables investigations considering the visual impression of the real room environment. Fig. 3.5 shows the HTC Vive with controllers and the lighthouse base stations used for tracking the position and orientation of the listener's head.

Figure 3.5.: HTC Vive (First generation) with controllers and Lighthouse base stations (1.0) for tracking (picture taken from [235]) and Vive Tracker (2.0).

Visual environment shown in the HMDs Throughout the studies, the participants had to wear one of these head-mounted displays (HMDs) which provided a neutral visual environment. This was advantageous because seeing the actual room is known to affect the perception of the binaural reproduction, particularly of the room acoustics in the auralization. Fig. 3.6 shows a screenshot of this neutral grid environment, which is a default environment provided by the system. On the floor, a polar grid indicates the original orientation of the system's coordinate system, determined by the room scaling procedure. In addition, one line is marked with an arrow. This line was used as a translation line in the corresponding experiments.

The HTC Vive system provides a mechanism to ensure that the user remains within that area and does not hit any obstacles around. If the user gets to close to the boundaries, an obvious blue grid fence visualizes the range of the area of activity.

3.4.2. Choice of headphones and headphone compensation

For the earlier experiments within this thesis, the participants had to wear an HMD. In these cases, STAX SR-202 headphones driven by a SRM-252 II amplifier unit (STAX SRS-2050) were used for the audio reproduction.

Headphone transfer characteristics are known to be a major source of spectral coloration. Therefore, a compensation filter is of interest. Lindau and Brinkmann [54] showed that for the auralization of non-individual binaural recordings, the headphone transfer functions (HpTFs) measured with that same HATS should be used for headphone compensation. Non-individual HpTFs were determined by measurements with the STAX headphones placed and repeatedly replaced on the ears of the

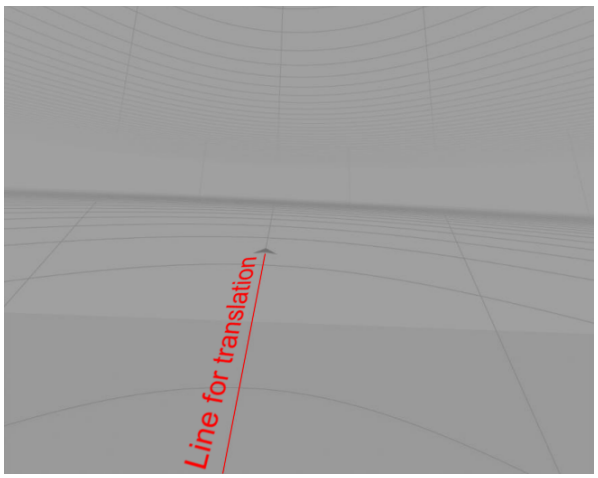


Figure 3.6.: Neutral grid environment which is shown in the HMDs during the listening session in the studies. It facilitates the orientation within the virtual environment without providing any information about the acoustic scene.

Kemar 45BA. For the creation of the headphone compensation filters, the HpTF were inversed with the least squares approach as described by Schärer and Lindau [53] with roll-off frequencies of 80 Hz and 20 kHz.

The last two experiments in this thesis consider an AR scenario. The listener does not have to wear an HMD and, thus, can see the real environment. The listener's motion is captured with an HTC Vive Tracker attached to the headphones. In the context of Augmented Reality, it is of interest that the listener can hear his real environment well, and the presence of the interference with that caused by the presence of the headphones should be minimal.

Therefore, extra-aural headphone models are interesting for such applications. These can be placed at a short distance from the ears without fully covering them. The model BK211 was developed for research purposes [210]. However, their presence still has an effect on the sound field close to the listener's ears. Schneiderwind et al. [278] analyzed the physically measurable and audible effects caused by the presence of headphones, considering several open and three extra-aural headphone models. The BK211 caused the least disturbance of the sound field, and the audible consequences were minimal. Their construction is quite heavy and unstable on the head. That is a disadvantage for investigation with interactive position changes. If a listener is about to explore a sound field by walking around, a heavy and unstable headphone may influence his natural motion. Therefore, for the experiments aiming at AR scenarios, the extra-aural headphone model AKG K1000 was chosen. This model had the second least effects on the listener's perception of the actual environment.

3.4.3. Rendering for position dynamic binaural audio reproduction

For the main investigations of this thesis, the developed tool pyBinSim realizes the partitioned convolution and organizes the dynamic selection of BRIR filters. It is used with block sizes of 256 or 512 samples at a sampling rate of 48 kHz, depending on the calculating power of the computer, the reverberation time of the room the BRIRs were measured in and consequently, the filter length, which had to be chosen in correspondence. In all experiments within this thesis, never more than one virtual sound source was active at a time. Headphone compensation filters were used for all experiments.

In order to keep the auralization of the created BRIR data as pure as possible, no position-

dependent interpolation was used for the reproduction. Only a short cross-fade was applied in the time domain to avoid switching artifacts when adapting the BRIR filter.

Test signal This thesis conducted most experiments with dry male speech as a test signal. It was taken from the audiobook (Arthur C. Doyle's "Sherlock Holmes"). Fig. 3.7 visualizes the spectral properties.

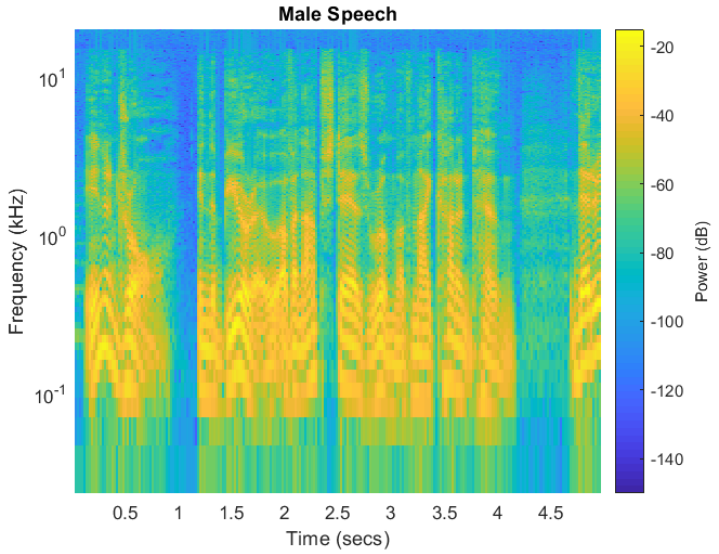


Figure 3.7.: Spectrogram of a 5-second excerpt of the dry male speech test stimulus used for the experiments within this thesis. The male voice is reading an audiobook. For the investigations, excerpts with a length ranging between 2 and 8 min excerpt were used.

3.5. Initial Study - First realization and informal attempt to evaluate plausibility

This section describes the investigation described in the paper "Binaural walk-through scenarios with actual self-walking using an HTC Vive" that I wrote together with Niklas Knoop. It was presented in 2017 at the Annual German Conference on Acoustics (DAGA) in Kiel, Germany.

The sensitivity to deviations in the sound field progress can be measured in various ways. The studies presented in this thesis are built on the idea that the main goal is the creation of convincing auditory illusions. Criteria to measure the quality of an auditory illusion have been discussed in Sec. 1.5. An ultimate target would be the creation of an illusion that is perceptually identical to its real counterpart and, thus, considered *authentic* [29]. However, with the early implementations of position-dynamic binaural reproduction realized in this thesis, the first goal is to achieve *plausibility*.

From their experiences in everyday listening, people, to a certain extent, have an idea of what it sounds like walking toward or past a sound source. Therefore, they have certain expectations towards a virtual version. This is referred to as the internal reference. Plausibility describes the agreement of the heard scene with an internal reference [236]. Plausibility is achieved if the illusion fulfills the listener's individual expectations and, thus, is perceived in agreement with the listener's internal reference.

Lindau and Weinzierl [211] proposed a method to evaluate the plausibility of a system for dynamic binaural synthesis. It was shown that most participants could not confidently identify the binaural

auralization among the real versions of the studied binaural reproduction system. The presentation of a real scene within the test items might influence the internal reference. Thus, the requirements to achieve plausibility might differ in a test design without a real scene.

A first experiment [321] was conducted to validate the approach of synthesizing a dynamic binaural reproduction for interactive listener translation along the given line with the measured BRIRs using only a very short cross-fade in the time domain, but no interpolation.

The test included the presented frontal source and an additional sound source on the right side of the line. Dry male speech and music (pop song - Michael Jackson "Give into me") were used as test stimuli. For the frontal sound source, all participants rated the created scene as plausible in a Yes/No test paradigm. For the side source, most participants were still happy with the reproduction. Fig. 3.8 shows the reproduction setup used in the experiment.

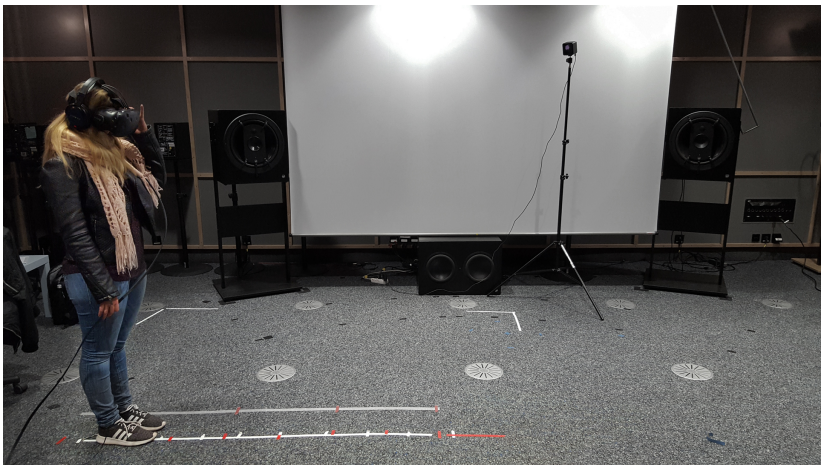


Figure 3.8.: Setup for the dynamic headphone reproduction during the experiment. The participants had to wear an HTC Vive HMD for the tracking and had to walk along the given translation line, which could be seen in the virtual environment as well [237].

Lindau and Weinzierl [211] argued that rating plausibility on a linear scale between 0 and 1 would be influenced by a strong bias due to individual theories on the plausibility of spatial auditory illusions, particularly if no reference is presented.

This was confirmed by the results of the first experiment described by Neidhardt and Knoop [321]. The participants had to rate plausibility on a scale from 1 to 100. The results showed a nearly equal distribution of the ratings over the whole scale range.

In a second attempt, participants were asked to rate plausibility in a Yes/No paradigm without a real scene among the test items. Although the scenes were presented in a randomized order, all participants rated the BRIR set measured with the loudspeaker in the front as plausible for both signals, dry male speech, and pop music. However, participants often mentioned that a scene was either clearly plausible or, in other cases, they found it hard to decide between Yes and No. Hence, it may be helpful to distinguish between "clearly yes" and "rather yes" as well as "rather no" and "clearly no" to capture such tendencies. In [321], the participants were additionally asked to rate the impression of walking toward a sound source. The ratings are expected to be highly correlated to plausibility judgments because the impression of walking toward the source is essential for plausibility. The considered Yes/No ratings were nearly identical to the plausibility ratings.

Besides the interactive binaural audio scenes created from the measured BRIR data set, scenes based on simulated BRIRs with two different sets of absorption coefficients (Simulated 'Lab' with $T_{30} \approx 0.2\,s$ and a more reverberant room with $T_{30} \approx 1.1\,s$) realized with MCRoomSim [233] were included in the evaluation.

Results and observations The two columns in Fig. 3.9 give an overview of the answers to both questions for the different test items. Although all scenes were evaluated in a randomized order, the scenes which differ only by the source signal show similar results. A clear preference for the scenes based on measured BRIRs can be observed. A longer T_{30} seems to reduce the drawbacks of the simulated BRIRs.

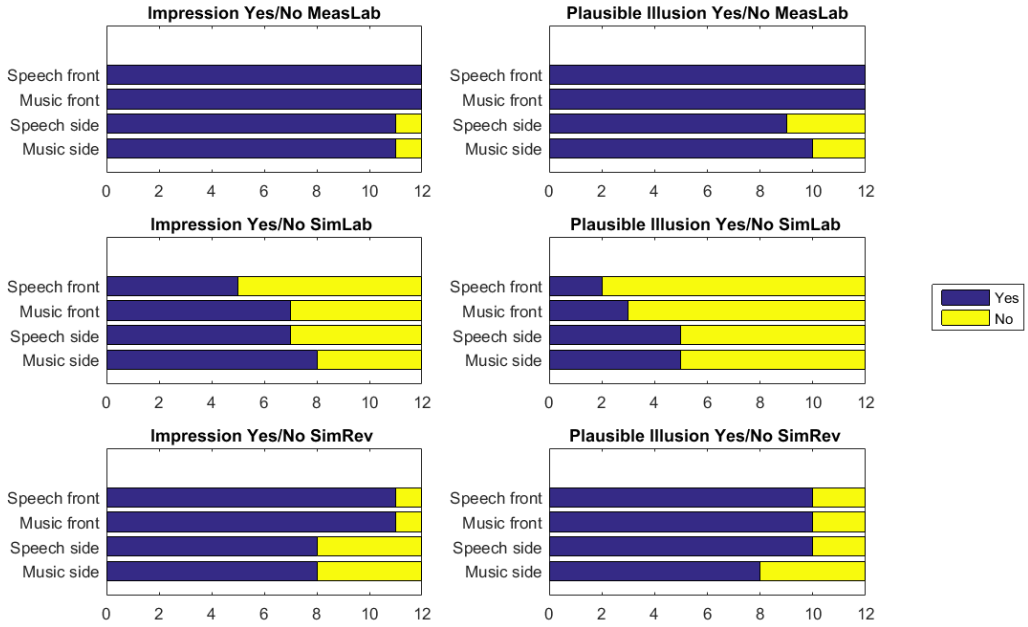


Figure 3.9.: Results of second experiment based on Yes/No questions: Answers to "Did you get the impression of walking towards/past a sound source?" (left), Answers to "Would you call this experience a plausible illusion of a sound source?" (right)

The dryer room was rated plausible by considerably fewer participants than the measured scene, but the more reverberant simulation was nearly as good. For this exploratory study, with 12 participants rating each scene once, a detailed statistical analysis did not seem helpful. Overall, the results suggested that the realized setup for interactive listener translation could create a plausible illusion of walking toward and past a virtual sound source based on the measured BRIR data set. With simulated BRIRs, the plausibility seemed to be influenced by the reverberation time of the simulated room.

4. Perception of room acoustics during continuous change of listening position

The previous chapter described the implementation of a system for position dynamic binaural synthesis and a first informal evaluation of the plausibility of the induced auditory illusion. The results of this informal pilot study indicated that plausibility was achieved with the measured BRIR set for most of the participants.

This chapter presents the main investigations on the influence of impaired or simplified room acoustics on the plausibility of walking towards, past, and away from a virtual sound source with the realized system. In addition to the BRIR data set presented in the previous chapter, the same measurement setup was moved to a substantially more reverberant seminar room. Moreover, the loudspeakers were turned by 180°, facing away from the translation line. This creates the case of the listener moving behind the sound source.

First, two studies (4.1 and 4.2) are presented that explore the impact of different BRIR set impairments on the plausibility of walking towards and away from a virtual sound source in the two different rooms. Some simplifications were chosen to be quite severe to explore the boundaries of plausibility. Both experiments were conducted with the participant wearing an HTC Vive HMD showing a neutral grid environment to minimize the impact of visual information on the evaluation.

The third study (4.3) examines the required resolution of the BRIR position grid with the given system. The fourth study (4.4) evaluates the plausibility that can be achieved with the given system in a Yes/No paradigm with and without a real hidden reference in an Augmented Reality scenario. The room environment, as well as the real loudspeaker that is supposed to emit sound virtually, can be seen by the participant throughout the test procedure.

The final study is based on a very similar AR scenario in the seminar room. Only in addition, also the "indirect irradiation case" with the loudspeakers turned around was taken into account.

Except for the BRIR grid study, all investigations are based on BRIR measurements conducted in the two real rooms, with their interior providing a complex geometric environment. Therefore, the test scenarios exhibit high ecological validity, particularly in the combination of the direct and indirect irradiation scenarios. This is in contrast to many studies conducted based on room acoustic simulations with an idealized environment and listening positions in rather advantageous locations like the sweet area. In a position dynamic exploration of the acoustic scene, the listener will likely leave these optimal positions. The inclusion of the indirect irradiation scenario gives credit to this.

The present chapter attempts to answer the following questions:
How do simplifications in the progress of the sound field affect perception?
Which simplifications can be applied with no or minimal consequences?
Which simplifications should be avoided because the quality of the auditory illusion would be affected too much?

The work on this thesis started in the summer of 2016. From this point until the day of the submission, at the end of 2022, there was considerable progress in the field of position-dynamic binaural room auralization. Therefore, this chapter starts describing the previous work available at the beginning of this work because that helps understanding why the experiment were designed in that way. The text is written in the order in which the studies were conducted to give an insight into the step-by-step process of gaining knowledge through one experiment and designing the next one in accordance.

4.1. Experiment I: Plausibility of walking towards a virtual sound source - in a listening lab

This section is based on the conference paper with full-paper review "Plausibility of an Interactive Approaching Motion towards a Virtual Sound Source Based on Simplified BRIR Sets" [81] that I published together with Anson Davis Pereppadan and Alby Ignatious Tommy at the 144th AES Convention in Milan, Italy, in 2018.

When a listener walks towards a sound source, the relative distance to the sound source changes. To create a plausible illusion of walking towards a virtual sound source, the distance change has to be modeled adequately. Therefore, it is vital to be aware of the parameters relevant to auditory distance estimation. Sec. 2.1 provides a short overview. As discussed in the introduction (Sec. 1.1), instead of recreating a perfect physical replicate of the sound field, a psychoacoustic optimization is of interest to keep the computation costs as low as possible. Suitable approximations of the parameters relevant to auditory distance estimation, such as the DRR or the change of sound level, are required. While walking towards a sound source, a listener changes not only his position relative to the sound source but also within the room. Thus, in the real sound field, the spatiotemporal pattern of early reflections changes as well. Also, the ITDG is likely to vary.

Wefers et al. [238], for example, presented a system for rendering virtual acoustic environments in real-time for binaural reproduction over loudspeakers in a VR-CAVE. The system was based on generating an accurate 3D model of the desired room and its acoustical properties. The room impulse responses are pre-calculated with simulations based on geometrical acoustics methods. In the later reproduction, they are convolved with the dry mono source signals. A quick system response to listener motion is required. For this reason, filters are updated section-wise. The direct sound is updated if the source or listener changes their position by a few centimeters or changes their orientation. In contrast, late reflections are updated only for position changes of at least 1 m.
For early reflections, it would be of interest to separately render the delays and properties of each individual image source up to a particular order to achieve high physical accuracy. However, this would increase the computational effort by several magnitudes and, thus, limit the possible number of virtual sound sources and simultaneous listeners in the system. The question about the perceptual relevance of these detailed properties could not be answered at that point.

Clapp and Seeber [239] investigated the effect of partially updated room auralizations in the case of source movement and observed an impact on the localization. The experiment was not conducted with binaural technology but with a 96-loudspeaker-circle following the nearest neighbor approach.

Boerum et al. [189] investigated binaural cross-fading to realize a smooth transition between BRIRs measured for positions that were 1 m apart. For the psychoacoustic evaluation, the participants were asked whether they had the impression of self-motion while listening to a continuous change in the observer position. The subjects did not physically change their position during the experiment. However, an interaction between physical self-motion and the expected change of the sound reaching the ears is likely and should be considered in such investigations.

Franck [240] proposed cross-fading in the frequency domain as a suitable method for dynamic binaural rendering. A perceptual comparison in an auralization considering interactive position changes is of interest. Mittag et al. [241] synthesized BRIRs from BRIRs measured at a few positions with different approaches. A validation concerning interactive walk-throughs was still open.

Pörschmann et al. [63] proposed the creation of BRIRs based on parameters extracted from a measured omnidirectional room impulse response. The authors claim that an authentic, indistinguishable reproduction is often not required, and for many applications, the creation of a plausible

scene is sufficient. By the date of this study, the plausibility of the synthesized BRIRs had been evaluated for head rotation, but not for interactive position changes of the listener. Plausibility for a static listening position might be easier to achieve than for translation. Pörschmann and Stade [242] also showed that the perceived auditory distance could be manipulated with the proposed approach. An evaluation with an interactive listener translation is of interest.

Generally, parameterization is of interest to reduce not only the measurement effort but also to reduce the required calculation power and memory space. In this context, the perceptual mixing time was proposed and studied [150, 195]. It varies with the properties of the room. The exact dependencies are not known yet. In theory, the mixing time approach should be applicable to position changes as well, at least for frequencies above the Schroeder frequency. However, validation is missing. The interaction of different auralization methods for position dynamic reproduction and the actual self-motion of the listener still needs to be studied in depth. Especially the quantification of thresholds is of interest. The sensitivity to changes in the room acoustics is relevant in this context.

The study presented in this section investigates the effect of various simplifications of the measured BRIR data set on the perceived plausibility.

Goal of the study The psychoacoustic requirements to achieve plausible illusions of walking toward a virtual sound source are not fully understood. Experiments that take interactive listener translation into account are important for profound validation.
On the one hand, this study will contribute to exploring potential test methods. On the other hand, an example set of measured BRIRs will be analyzed in order to identify potentials for inaudible data reduction. This experiment focuses on the role of room acoustics and the related psychoacoustic requirements of an interactive binaural walk-through. Is the listener sensitive to impairments in reverberation as they occur in interpolation and extrapolation methods? Will impairments like partly updated BRIRs affect the plausibility of an interactive approaching motion?

The auralization of the given BRIR set was rated as plausible by all participants in a Yes/No paradigm in the previous study, not considering a real scene. The presented experiment aims at determining the effects of various systematic simplifications of these BRIRs.

4.1.1. Methodology

Manipulation of BRIR data

A systematic simplification of the data is of interest because it can help to find approaches for data reduction. For the first test case, the positional resolution was reduced from 25 cm to 50 cm. Furthermore, this study focuses on a selection of physical parameters, which successively change along the line. For an investigation of their perceptual relevance, test scenes were created, in which selected parameters were kept constant.

One of those parameters is the ratio of direct to reverberant energy. To create the corresponding manipulated scene, the BRIRs measured at one position were used to create the BRIRs for the other positions by adjusting the level of the whole BRIR according to the attenuation of the original direct sound. This manipulation is, e.g., interesting with regard to inter-positional cross-fading as discussed by Boerum et al. [189]. The level of the reflection and, consequently, the reverberant energy changes inappropriately in this case.

A further interesting parameter is the predelay. It is defined by the duration of the direct sound traveling from the source to the ears and varies with the distance. In static binaural reproduction, a change of the predelay is usually not audible. In contrast, especially when walking toward a sounds source, the predelay changes continuously. Thus it becomes interesting whether a constant predelay in the reproduction of relative distance changes causes audible effects and influences the plausibility.

The initial time delay gap (ITDG) describes the time that passes between the arrival of the direct sound and the first reflection at the ears. The ITDG varies with the listening position and plays a role in the perception of the room and the distance to the source. If no other reflecting surface is nearby, the reflection on the floor or ceiling arrives first. In the present case, the room height is 2.8 m, the height of the ears at 1.59 m, and the center of the loudspeaker at 1.55 m. Thus, the ITDG will change from about 4.5 ms at the closest position to 2.3 ms at the farthest. Fig. 4.1 shows both BRIRs (left ear) separately.

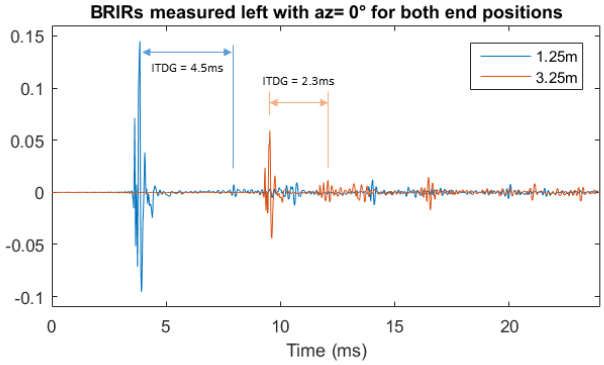


Figure 4.1.: BRIRs measured at the left ear for the closest and the farthest position.

If a source would be placed in an anechoic chamber, no reverberation would be available. Therefore, it is of interest, whether only the originally measured direct sound would create a plausible illusion of a sound source one can walk to. The original direct sound was separated from the room reflections.

The temporal and spatial structure of all reflections changes with the position as well. It has been shown in the past that the sensitivity to changes in the reflection pattern is limited [127], [323]. However, a lack of adaptation might affect the plausibility of the walk-through.

Hence, it was included in the test using the reverberation pattern of one position for all positions. For several manipulation cases, direct sound and reverberant part of the impulse response had to be separated. This was done using a half hann window (length of 32 samples) as fade-out of the direct sound part and fade-in of the reverberant sound part. The center of both half windows crossed 1.5 ms after the direct sound.

The following BRIR sets were included in the test:

B1 Original measured data
The complete set of BRIRs measured at 9 positions with distances of 25 cm serves as a reference in the test. The same scene had been rated as plausible in a Yes-No paradigm by all participants of the previous study [321].

B2 Original measured data - 50 cm steps
BRIRs from every second position were included in the reproduction, five positions in total

B3 Created from BRIRs of closest position - Adjustment of sound level and predelay
DRR, ITDG, and reflection pattern remain constant

B4 Created from BRIRs of farthest position - Adjustment of sound level and predelay
DRR, ITDG, and reflection pattern remain constant; DRR is lower than in B3

B5 Created from BRIRs of closest position - Only adjustment of sound level to match direct energy with original direct energy for the single positions
Predelay, ITDG, DRR, and reflection patterns remain constant.

B6 Original direct sound - no reverberation
Predelay changes appropriately (original values)

B7 Original direct sound - constant reverberation only from the closest position
ITDG and reflection pattern remains constant; predelay and DRR change appropriately

B8 Direct sound only from the closest position, level, and predelay of direct sound adjusted - constant reverberation only from the closest position
ITDG and reflection pattern remains constant; predelay and DRR change appropriately

Test setup

The experiment was carried out in the listening laboratory at the same positions where the measurements had been conducted. The participants had to wear an HTC Vive HMD. The Vive base stations tracked the position and orientation of the user's head throughout the experiment. The real-time convolution for the dynamic reproduction was realized with pyBinSim [317]. The Python tool selected the BRIR filters according to the current position and orientation of the listener. When switching between filters, a cross-fade shorter than the length of the block size (256 samples with fs = 48 kHz) was applied. No further interpolation was applied.

The line for listener translation with the length of 2 m was marked on the floor of the real room, and it was also marked in the virtual environment shown in the HMD. The visual scene was a neutral grid which helped the participants orient themselves but did not provide any visual cues with respect to the sound source or the reproduced room. The participants were asked to stay on the marked line because only there, corresponding BRIR data was available. Rotation in all (azimuth-) directions and with arbitrary speed was allowed. The audio was played back via STAX SR 202 headphones, driven by an SRM-252II amplifier. A headphone equalization filter determined from a non-individual measurement with the KEMAR 45BA was applied for compensation.

Test method

The participants were asked to walk up and down the line at least once. Thereafter they could explore the virtual scene arbitrarily and as long as desired. When the participant felt ready, the conductor switched off the audio reproduction and started to ask the following four questions with four given answering choices each.

Then four questions with four answering options were asked:

How would you describe the externalization?
1 - Perfectly external
2 - Mainly external with small constraints
3 - External with larger constraints
4 - Not external at all

Did you have an impression of walking towards/away from the source?
1 - Clearly yes
2 - Rather yes
3 - Rather no
4 - Clearly no

Did you perceive a continuous change of the sound field?
1 - Perfectly continuous
2 - Slightly noticeable steps
3 - Stronger steps
4 - Large annoying jumps

Would you call it a plausible illusion of a loudspeaker standing in a room?
1 - Clearly yes
2 - Rather yes
3 - Rather no
4 - Clearly no

If they were unsure of the answer, the participants were encouraged to listen to the scene again. In case a participant could not match his perception to the given choice of answers, he/she was asked to describe the own impressions in own words, and the conductor wrote everything down. Additional comments were written down as well.

Sixteen test scenes (eight BRIR sets with two different stimuli) were presented in a randomized order. Before the actual test, the conductor presented two scene examples to the participant, the complete measured BRIR set (B1) and the original direct sound without reverberation (BRIR set B6) with speech to get familiar with the task and with possible differences.

Participants

Twenty-two subjects (4 females and 18 males) participated in the test. The average age of the group was 27.9 years, ranging from 22 to 38. Nine subjects were trained listeners in binaural audio. For the other panelists, it was mostly their first experience with VAEs reproduced via headphones. All reported to have normal hearing abilities and not to be aware of any confinements.

4.1.2. Results

Three naïve participants rated all the scenes with only the best ratings that were possible. Their results were removed from the analysis because the goal was to evaluate the audible effects of the simplification. The answers of these participants suggest they could not perceive any effects caused by the BRIR manipulation. Fig. 4.2 provides an overview of the ratings given by the remaining participants for the different attributes. The bars labeled with S1 to S8 show the answers for the cases in which the virtual loudspeaker reproduced dry male speech, N1 to N8 label the scenes with white noise as source signal.

The four options for answering each of the four questions given in Sec. 4.1.1 describe a continuous degradation from 1 to 4. Consequently, the choice can be regarded as a continuous scale, though the distances between two steps might not be precisely equidistant. To test for significant differences between the various conditions, the Wilcoxon signed-rank test was used. The p-values of interest are provided in the discussion of the separate cases in the following subsections. The pre-defined significance level of $\alpha = 0.05$ was adjusted for multiple comparisons using the modified False Detection Rate (mFDR) approach discussed in [243], because it is less conservative than the Bonferroni correction. In this case, seven test conditions were compared to the measured reference. the corresponding mFDR-corrected significance level is 0.0193. For a comparison of each condition with each, the significance level has to be adjusted to $\alpha_i = 0.0127$ for the resulting 28 comparisons.

Because a new test method was used in this experiment, it is of interest to discuss the correlation between the various attributes. For this, the values 1-4 of the participants' answers were used. The Pearson correlation coefficient of two random variables, A and B, with N scalar observations, is defined as follows.

$$\rho(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{A_i-\mu_A}{\sigma_A}\right)\left(\frac{B_i-\mu_B}{\sigma_B}\right). \tag{4.1}$$

μ_A, μ_B are the means of A and B and σ_A, σ_B the standard deviation. To test whether the plausibility correlates, e.g., with externalization, each participant's rating for each scene is analyzed pairwise. Because the signal had a strong influence on the results, in some cases, the correlation coefficient

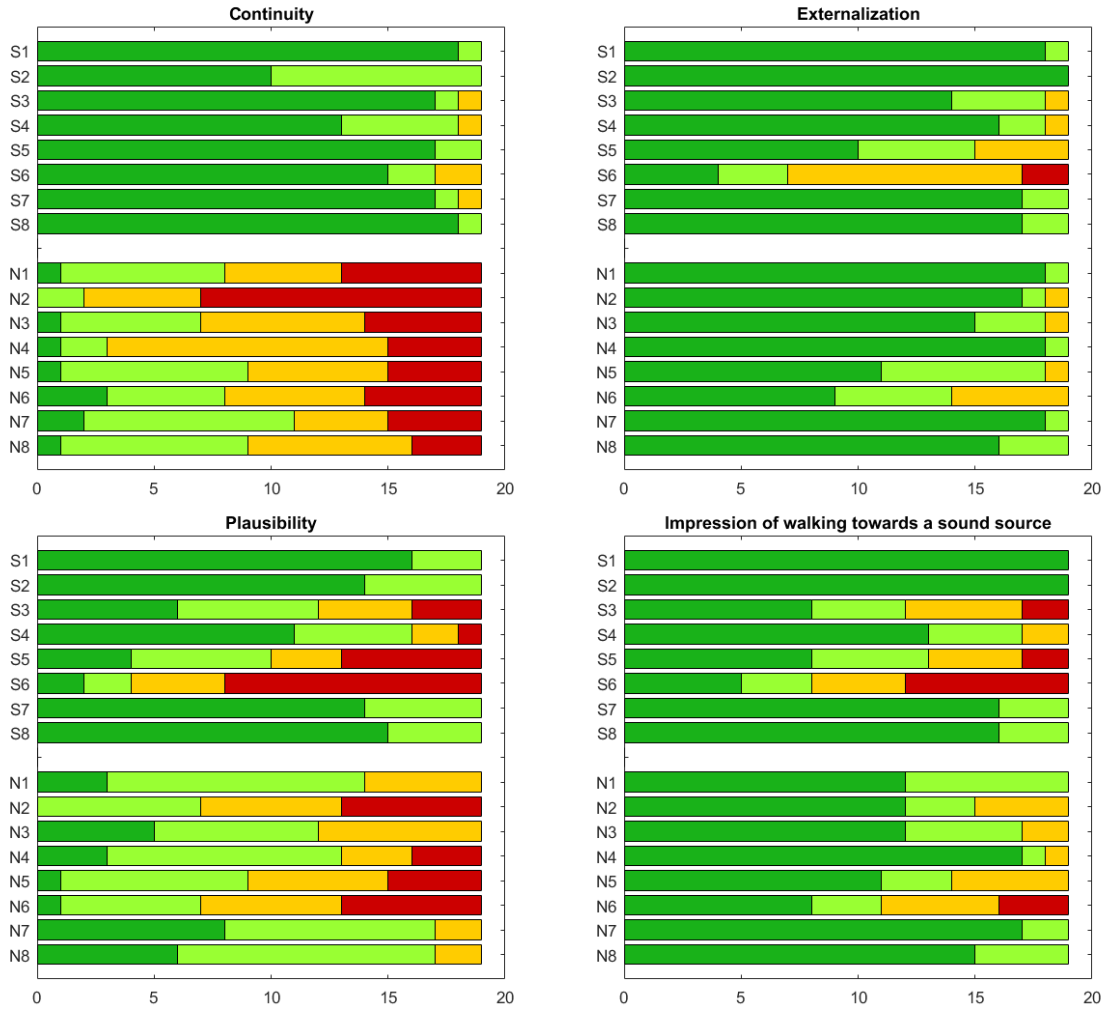


Figure 4.2.: Overview of the results - Continuity, Externalization, Plausibility and Impression of walking towards a sound source for the 8 different BRIR sets, S1-S8 with speech, N1-N8 with white noise. The bars show the numbers of participants who chose answer No. 1 (dark green), 2 (light green), 3 (orange) or 4 (red), as listed in Sec. 4.1.1.

was calculated separately for the speech scenes and the scenes with noise.

Continuity It is evident in the plot that the signal type plays an essential role in the judgment of continuity. With the speech, the reproduction was rated most often as continuous without noticeable steps. One reason might be that the translation line leads directly toward the loudspeaker. The relative angle between the listener and the sound source is not affected by the position change.
In contrast, with noise, the ratings on continuity were poor. Noise is generally a critical signal regarding spatial sampling. On the one hand, it is continuous in time, while speech often contains short breaks with silence. If filters are switched then, slight changes in coloration and source position are harder to notice. On the other hand, noise contains all frequencies, while male speech is limited to a low-frequency band. Higher frequencies are more critical to spatial sampling.
For BRIR set B2 with the 50 cm spacing between the BRIR positions, there is only a small but significant decrease in ratings compared to the full data set with 25 cm inter-positional distances for speech (p=0.019), but not for noise (p=0.028).

The perception of severe limitations in continuity affects plausibility. The plausibility ratings for the test items with noise are worse than for speech ($p < 0.0001$). However, the correlation between

continuity and plausibility is only 0.35 over all ratings.

Externalization The full data set as well as the set with the reduced positional resolution, were rated as perfectly external with speech. In a reproduction with constant reverberation, as with B7 and B8, the externalization is not affected ($p = 1.0$ for both scenes and signals). A significant decrease compared to the original BRIR set can only be found for B5 ($p = 0.007$ for speech and $p = 0.015$ for noise) and B6 ($p < 0.0001$ for speech and $p = 0.004$ for noise).
For B6, which is the scene produced from the original direct sound without room reflection, the externalization has been rated worse. It is commonly known that the lack of reverberation affects externalization significantly [244].
B5 is created only by changing the overall level of the whole BRIR. It lacks variations in DRR, predelay, and ITDG. B3 and B5 only differ by the adjustment of the predelay in B3, their externalization is not significantly different ($p = 0.2$). The externalization in B4 is significantly better than for B5 ($p = 0.009$).
B4 is built from the BRIRs measured at the farthest position. Thus, the DRR is lower than in B3 and B5 throughout the scene. This might support externalization.

The limited continuity of the scenes with white noise did not affect the externalization. In contrast, degraded externalization seems to be linked to decreased plausibility. The correlation coefficient of the results of the speech scenes for externalization and plausibility is 0.68 and 0.54 over all scenes. Of course, externalization is essential for plausibility. For the noise scenes, the correlation coefficient is 0.43. This might be due to a dominating influence of bad continuity.

Plausibility and impression of walking towards a source The question for the impression of walking towards the source was included in the test to confirm its connection to plausibility. A high correlation was expected because a plausible illusion of an approaching motion cannot occur without having the impression of walking toward the source. Fig. 4.2 shows the results. Especially in the case of the speech signal, the relation is evident in the results. The correlation coefficient is 0.84 for the speech scenes, 0.53 for noise, and 0.6 overall. However, the ratings regarding the impression of walking towards the source are less critical than for plausibility.
As already discussed, degradations in continuity and externalization affect plausibility. This emphasizes the role of plausibility to serve as an attribute for overall quality.

For the scenes with the speech signal, clear differences can be found in the ratings. The original full set of measurements was perceived as the most plausible. B7 with the original direct sound and the constant reverberation was rated equally well for speech ($p = 0.63$). The constant reverberation did not affect the plausibility. Furthermore, reasonable plausibility can be observed for B2 ($p = 1.0$) with the reduced spatial resolution and for B8 ($p = 1.0$), which was created from the BRIRs of one position by adjusting the level of the direct sound. Thus plausibility was not affected by replacing the original direct sound with one filter that is adjusted in level throughout the reproduction.

The worst rating was given for S6, and the BRIR set without reverberation (significantly different from S1 with $p = 0.007$). Externalization is affected by the lack of sound reflections, and so is plausibility. Some participants reported that the sound source followed when they tried to move away from it or "it was sitting in the neck," which was perceived as "creepy." This phenomenon can also be described as "an insufficient change over distance." The listeners expected a greater change due to their movements.
Although the reverberation is relatively low in the measured room, the absence of the reflections degrades the plausibility. This might be connected to the observation by Nielsen [50] regarding the auditory distance perception in an anechoic room.

In addition, the scenes S3 and S5 were rated significantly different from the reference S1 regarding the impression of walking toward a sound source and plausibility (all $p < 0.005$). In both scenes, the

DRR remains constant over the distance change, which is different from S7 and S8. Consequently, a lack of change in DRR affects the plausibility of walking towards a sound source. That in unsurprising because the DRR is known to be an important cue for distance perception - and thus, for the perception of the change of distance.

The ratings for S3 are not significantly different from S5 ($p = 0.18$). In B3, the predelay was adjusted to the different positions. Although no significant influence of the predelay could be found here, it cannot be concluded that it is not essential. The predelay becomes interesting with regard to the doppler effect. A lack of change in predelay might become perceivable with quicker changes of distance than in the case of a walking and carefully exploring listener.
S4 was rated slightly better than S3 ($p = 0.053$). The constant but lower DRR might be of advantage. However, none of the scenes with constant DRR is as plausible as the original measured data set.

For the scenes with white noise, the poor continuity seems to dominate the plausibility ratings and mask potential effects of the different BRIR manipulations. The trends are similar to that observed in the speech scenes, but they are less pronounced.

4.1.3. Discussion and Conclusion

Physical analysis

The DRR is discussed as an important cue in distance perception. However, the parameter provides only limited insight into the contribution of direct sound and reverberation to DRR changes between the positions. Therefore, additionally, the direct sound energy and the reverberant energy along the line were set in relation to the energy of the direct sound at the closest distance to the sound source.

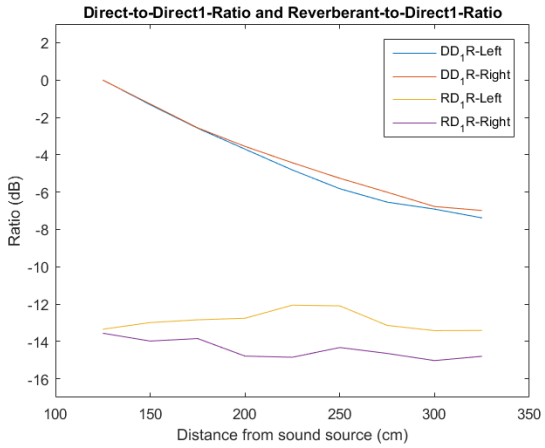


Figure 4.3.: Analysis of the contribution of direct and reverberant energy to variations in the DRR: Ratio of the direct sound energy and reverberant energy along the line to the direct sound energy of the closest position for an azimuth angle of 0°.

Fig. 4.3 visualizes the results for both ears at an azimuth of 0°. This allows for inspecting the relative variations of direct and reverberant energy along the translation path.
At the right ear, the reverberant energy shows a slight decrease with increasing distance. However, the variations remain within 3 dB and are decreasing not steadily but with fluctuations. Those might also be due to uncertainties in the measurements. For the left ear, the reverberant energy shows similar fluctuations but without a clear trend.

Thus, the decrease of the direct sound energy along the path dominates the relative changes of the DRR in the analyzed BRIR set. This fact might be essential with regard to the observation that the same reverberation filter could be used throughout the path.

Discussion of the method

The test design brought up distinct differences in the perceived plausibility and allowed for observing certain links between the different attributes. Thus, the test method appears to be appropriate. Plausibility is a suitable attribute for estimating the overall quality.

Due to the relatively long exploration time per scene, every participant rated each scene only once. Hence, conclusions on reliability and repeatability are difficult. The study presented in this section is still exploring ideas and approaches to investigate and compare the perception of interactive walk-throughs. The results will help design advanced test methods.

Summary and conclusions

The presented experiment investigated the influence of various simplifications of a measured BRIR dataset on the perceived plausibility of an interactive approaching motion toward a virtual loudspeaker. The BRIRs were measured along a line leading toward a loudspeaker in a real, relatively dry room. The measured data was systematically impaired to study how the separate physical parameters the plausibility of the scene.

White noise is a very critical signal with regard to a limited spatial resolution of the available filter set. The combination resulted in poor ratings of the continuity and affected plausibility. The spatial sampling of the direct sound is probably the dominating aspect. An investigation with a continuous adaptation of the direct sound is of interest to study the required resolution for the reverberation.

Although the reverberant energy is very low in the measured data, eliminating the reverberation leads to a significant reduction of plausibility. A constant spatiotemporal reflection pattern throughout the scene did not affect the plausibility of the given data set. The reverberant energy is generally very low in the chosen room. With room reflection masking in mind, these observations are not surprising. It is still possible, though, that stronger reverberation is even less critical with regard to changes in the early reflection pattern.

Furthermore, switching the filter section for the direct sound is not required. Only an adjustment of level and predelay seems to be essential to maintain plausibility.

The results suggest that a lack of change in DRR and predelay affects plausibility. This is interesting with regard to cross-fading of BRIRs in the time domain as proposed in [189]. A detailed investigation of this aspect is required.

Externalization was not affected by poor continuity, but as expected, the elimination of the reverberation degraded externalization. In addition, constant DRR seems to have an impact as well. However, the observed effect could also be due to the low level of reverberation. The differences between the scenes are still relatively small, and not every naïve listener can perceive them.

The measured line is entirely within the critical distance of the source. The direct energy is dominant compared to the reverberant energy. The experiment should be repeated with a data set measured in a more reverberant environment, either outside or at the critical distance.

4.2. Experiment II: Plausibility of walking towards a virtual sound source - in a seminar room

This section is based on the conference paper with full-paper review "Plausibility of an approaching motion towards a virtual sound source II: In a reverberant seminar room" [324] that I published together with Samaneh Kamandi at the 152nd AES Convention, Online, in 2022.

In the previous experiment, keeping the reverberant part of the BRIR constant throughout the potential listening positions did not affect the plausibility of the interactive approaching motion towards a virtual loudspeaker. The translation line was located within the critical distance of the sound source. Consequently, the direct sound energy was stronger than the energy of the reverberant sound throughout the whole scene. Furthermore, the listening laboratory chosen for the study exhibits a relatively low reverberation time. Therefore, it is of interest whether a test scenario with stronger reverberant sound energy in relation to the direct sound yields similar observations. On the one hand, in the case of stronger reverberation, its spatiotemporal structure may play a more critical role in the perception. On the other hand, an increase in the reverberant sound energy will be accompanied by an increase in the density of strong reflections. This may raise the perceived diffuseness and thus reduce position- and direction-dependent perceptual differences.

In order to assess the role of stronger reverberation in the perception of a manipulated and slightly simplified progress of the reflection pattern in an interactive approaching motion, a second experiment was conducted in a seminar room of the university in Ilmenau.

4.2.1. Measurement: BRIRs for translation in seminar room

In order to study whether the observations hold for rooms with stronger reverberation, a second data set was created with a similar measurement setup in a seminar room. The room has a size of 9.9 m×4.7 m×3.1 m, a volume of V = 144.3 m^3 and a reverberation time of T_{30} = 0.98 s (broadband).

Again, a line with a length of 2 m leading towards and past a Genelec 1030A loudspeaker was chosen. To achieve a similar direct sound, the range of the distances to the loudspeakers remained the same (1.25 m to 3.25 m to the center of the loudspeaker or 1.12 m to 3.12 m to the membrane).

Equivalently to the previous measurement, the Kemar 45BA was subsequently placed at 9 positions in intervals of 25 cm. Figure 4.5 illustrates the arrangement of the measurement positions in the seminar room. The ears were located 1.59 m above the floor as in the listening lab, and the center of the loudspeaker at height of 1.55 m. Consequently, its acoustic center was located approximately at the ear height. Again, BRIRs were measured with an angular resolution of 4° in the horizontal plane.

4.2.2. Methodology

The main idea of this second experiment is to find out whether the conclusions drawn from the previous experiment in the listening laboratory still hold in a more reverberant room. Thus, the first experiment is repeated in a very similar way. In addition to the BRIRs measured in the seminar room and the corresponding simplified versions, the original scene measured in the listening laboratory is included in the test to find out whether it is still perceived as plausible in the new environment.

Moreover, the results of the previous experiment indicate that the scenes with a constant reverberation over different positions are very interesting with regard to an efficient simplification of the rendering. Therefore, additional test cases were considered. In the previous experiment, B7 and B8 were the scenes with a constant reverberation from position 125, once with the originally measured direct sound (B7) and with a direct sound that was only adjusted in sound level and pre-delay (B8).

Figure 4.4.: Photos of the BRIR measurement in the seminar room.

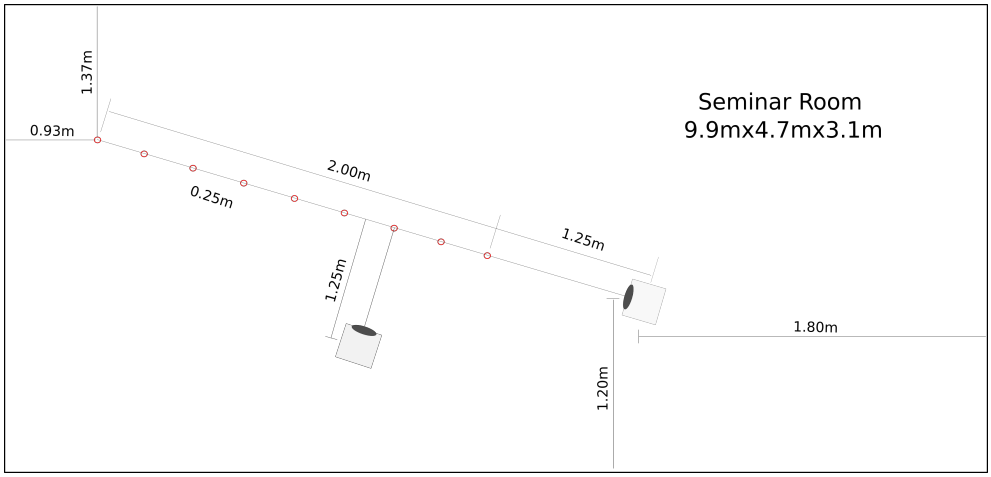


Figure 4.5.: Illustration of the source and measurement positions in the seminar room.

In this second study, equivalent test cases with a constant reverberation from position 325 were added (S9, S11, and S13). Furthermore, with S12 and S13, cases with time-aligned direct sound and constant reverberation were introduced.

Manipulation of the BRIR data set

The full BRIR set initially measured for those positions will serve as a reference scene in the experiment. The manipulations and simplifications are chosen in a way that challenges plausibility and explores its limits. Therefore, also some very keen simplifications were included. In addition, simplifications that are part of already proposed interpolation approaches are taken into account, like time-aligned direct sound or the manipulation of the Initial Time Delay Gap.
As a first step, the direct sound was separated from the reverberant part of the BRIR at 1.5 ms after the manually determined peak of the direct sound. A Hann window with a length of 64 samples was applied to both sides with a complete overlap. After manipulating the desired properties of direct sound and reverberant part, both parts were concatenated again with a complete overlap of both sides of the Hann window. The properties of interest are discussed below.

Pre-delay The pre-delay corresponds to the duration the sound needs to travel from the sound source to the listening position. Arend et al. [85] proposed to neglect it and work with time-aligned direct sound.

Sound level and DRR The level of the direct sound and its relation to room reflections changes with the listening position. This relation is perceptually relevant.

Initial Time Delay Gap - The Initial Time Delay Gap (ITDG) in a BRIR is defined as the duration from the arrival of the direct sound till the arrival of the first reflection. An adjustment of the ITDG is a fundamental part of the method proposed by Werner et al. [191].

Spatio-temporal pattern of early reflections - The times- and directions of arrival of all the early reflections vary with the listening position and may be perceptually relevant.

The following test scenes were created:

S1 - Originally measured data set - The measured BRIR set is the reference scene. All parameters vary according to their progress in the original sound field.

S2 - BRIR data set measured in listening lab In the previous study [81], the BRIR data set measured in a listening lab was rated "clearly plausible." The progress of the direct sound is similar to the seminar room scenario.

S3 - Constant BRIR from 1.25 m, only adjusted in sound level The BRIRs from the closest position are adjusted only in sound level. The level of the whole BRIR is adjusted according to the energy in direct sound with the distance. pre-delay, DRR, and ITDG remain constant over the positions.

S4 - Constant BRIR from 3.25 m, only adjusted in sound level The BRIRs from the last position are adjusted in sound level. The level of the whole BRIR is adjusted according to the energy in direct sound with the distance. pre-delay, DRR, and ITDG remain constant over the positions.

S5 - Created from BRIRs at 1.25 m Adjustment of sound level and pre-delay to match energy and time of arrival of manipulated and measured direct sound at each position DRR, ITDG, and reflection pattern remain constant.

S6 - Created from BRIRs at 3.25 m Adjustment of sound level and pre-delay to match energy and time of arrival of manipulated and measured direct sound at each position. DRR, ITDG, and reflection patterns remain constant.

S7 - Original direct sound without reverberation The original pre-delay is preserved. Without reflections, there is no ITDG or DRR.

S8 - Original direct sound and constant reflection pattern from 1.25 m - The original pre-delay is preserved. The DRR changes with distance, but not necessarily in the same way as in the original BRIRs. The ITDG remains constant.

S9 - Original direct sound and constant reflection pattern from 3.25 m - The original pre-delay is preserved. The DRR changes with distance, but not necessarily in the same way as in the original BRIRs. The ITDG remains constant.

S10 - BRIR from 1.25 m, only adjusted in the level of direct sound and pre-delay - The original pre-delay is preserved. DRR changes with distance, but not necessarily in the same way as in the original BRIRs. The ITDG remains constant.

S11 - BRIR from 3.25 m, only adjusted in level of direct sound and pre-delay - The original pre-delay is preserved. DRR changes with distance, but not necessarily in the same way as in the original BRIRs. The ITDG remains constant.

S12 - BRIR from 1.25 m, only adjusted in level of direct sound ITDG is constant. DRR changes with distance, but not necessarily in the same way as in the original BRIRs. The ITDG remains constant.

S13 - BRIR from 3.25 m, only adjusted in level of direct sound ITDG is constant. DRR changes with distance, but not necessarily in the same way as in the original BRIRs. The ITDG remains constant.

Test setup

The experiment took place at exactly the same positions where the measurements were conducted earlier on. In accordance with the first study, the participants had to wear an HTC Vive head-mounted display (HMD) which was tracked in position and orientation. The audio rendering software pyBinSim [317] selected the BRIR filter according to the tracking data and convolved it with the dry source signal. The partitioned convolution was realized with a blocksize of 256 samples at a sampling rate of 48 kHz. When the listener passes the center between two neighbored measurement positions, pyBinSim switches the filters. A short cross-fade over time eliminated switching artifacts. No further interpolation was applied.

In the HMD, a neutral grid environment was shown. It supports a general orientation in the scene but does not provide visual cues with regard to the source position or the audible room. It also isolated the participants from visual cues of the real environment, but they might still have had an image of it in their minds because they saw (and heard) the real room before the test.
The audio was reproduced with STAX SR 202 headphones driven by an SRM-252II amplifier unit. To minimize the influence of the headphones, a non-individual compensation filter was created with the least-squares approach described by Schaerer et al. [53].

The same eight minutes long excerpt of dry male speech reading an audiobook was used as the source signal as in the previous study. The excerpt was looped, and the participants could listen to the scene as long as desired. In the previous experiment, white noise was used as an additional test stimulus, but in that case, the perceived continuity was not sufficient. A substantially higher positional resolution would be required to study similar effects with white noise. Therefore, this experiment focuses on the given speech signal.

Test method

Similar to the previous experiment, this study explores the perceptual effect of a physically simplified sound field progress during listener translation on plausibility, sound source stability, externalization and continuity of the interactive approaching motion. Furthermore, this method to evaluate plausibility without considering a real reference is subject to exploration.
In order to acquire results which can be compared to the previous study, the same test method with the same attributes and questions was chosen. Although the perceived continuity was not the subject to research in this experiment, it was captured to ensure the resolution of 25 cm was adequate for the new test cases as well. Moreover, some of the approaches to create a simplified BRIR set could affect the perceived continuity. For this reason, the perceived continuity was captured.

During that earlier experiment, sound source instabilities were noticed in some cases, but this was not covered by the evaluation. Therefore, a question for sound source stability was added to the questionnaire:

Does the sound source move while you are moving?
1 - Clearly no
2 - Rather no
3 - Rather yes
4 - Clearly yes

For each scene, the participant had to walk up and down the full translation line at least once before rating. Additionally, he/she could explore the scene with arbitrary self-motion along the given line and with rotation. When the participant felt ready, the test conductor switched off the audio reproduction and asked the questions. If a participant felt like listening to the current test scene again in order to pick the best matching answering option, further exploration phases were allowed. When all questions were answered, the exploration phase for the next test scene started.

This procedure is closely related to the Absolute Category Rating described in ITU-R BS.1284-2 [245] .

Before the main test, the participants had to undergo the short training. This training consisted of three test scenes. The first was the BRIR set initially measured in the given room (S1) because it was expected to be the test scene with the best perceptual results. The scene without any reverberation (S7) was chosen as the second scene for the training because the perceptual difference to the first scene is evident and noticeable even for inexperienced listeners. The third scene for the training was S4 because it was neither of obviously good nor lousy quality. The procedure in training was the same as in the actual test in order to prepare the participant for the upcoming questions and to ensure that the task was well-understood before beginning the main experiment. Since the scenes from the training were also part of the experiment, no feedback was given to the subject with regard to their answers in training. They were only told that the scenes in the training cover the range of quality that can be expected.

In contrast to the previous experiment, each scene was tested twice in this second study. The order of all test scenes was randomized individually for each subject. The training and test sessions took between 40 and 70 minutes in total. For inexperienced listeners, this was slightly too long. They were encouraged to take breaks, and in a few cases, the test was continued another day.

Participants

Twenty subjects (11 female, 9 male) with an average age of 26.1 years with a standard deviation of ± 3.9 years participated in the test. All subjects stated to have no known hearing impairments. Seven out of 20 participants were experienced in evaluating binaural walk-through scenarios. Two of those experienced panelists had listened to the test scenes prior to the experiment. The other 13 subjects have not experienced dynamic binaural synthesis more than once before.

4.2.3. Results

Each of the 20 participants rated each scene twice for each of the five attributes. This adds up to a total of 40 ratings per scene and attribute. Fig. 4.7 provides an overview. Moreover, the repetition enables analyzing the consistency of each subject's ratings.

Consistency of the ratings

Sec. 4.2.2 discussed the challenges of the chosen method to evaluate plausibility. Therefore, it seemed useful to provide a detailed analysis of the rating consistency. The first graph in Fig. 4.6 shows the number of repetitions, wherein the ratings varied between the first and the second evaluation of the same scene per candidate over all attributes. The seven experienced listeners are labeled with 'T'. The graph visualizes how often a difference of one, two, or three steps on the rating scale occurred. A difference of one was observed quite often but is not considered critically inconsistent. In several cases, the participants found it hard to decide between two of the answering options and might have chosen the other of both options in the second trial.

In contrast, a deviation of two is a considerable difference, and three is the maximum possible. Both cases should not occur. Especially for the inexperienced listeners, it could be observed that they developed a better understanding of what to listen for during the test. Consequently, in the second part, when they rated each scene again, they made better use of the scales. Furthermore, the speech signal chosen for the test also contained regular short breaks, which may temporarily hide some drawbacks from the subject and may explain some of the occurring inconsistencies.

The results of the trained listeners T1 to T7 indicate that achieving a consistent rating was not easy. It is required to explore the whole scene in detail in order to be able to reproduce the

own ratings. Although encouraged to do so, several participants were not entirely consistent with this procedure. This is probably a matter of specific training and experience with position-dynamic binaural synthesis.
The ratings of participants 2 and 4 contained the highest absolute number of inconsistent ratings as well as the highest number of maximum differences between the first and the second rating. Therefore, their results were not considered any further. Interestingly, for P2, the four cases of the maximum difference occurred only for scene S4, and for P4, all cases of maximum difference occurred only for scene S2.
Furthermore, it was interesting to notice that a listener who had no experience with listening to dynamic binaural reproduction achieved the highest consistency while still making good use of the full scale. In addition, it is interesting that this participant did not complete the test in one session but finished the second part one week later and still achieved such a high consistency.
The second graph in Fig. 4.6 visualizes the distribution of inconsistencies over the different test scenes and attributes in the order of Continuity, Externalization, Stability, Impression of walking towards a sound source, and Plausibility. The graph indicates that the inconsistencies are not focused on a particular attribute or scene.
Still, some variations between the scenes can be found. The highest confidence was achieved for scene S10, followed by S1 and S1s, which are the reference and the two scenes constructed from the BRIRs of one position and adjusting only the level and predelay of the direct sound. In contrast, S4, S2, and S12 were scenes with the most considerable deviations, mainly caused by participants 2 and 4, whose results were excluded afterward.

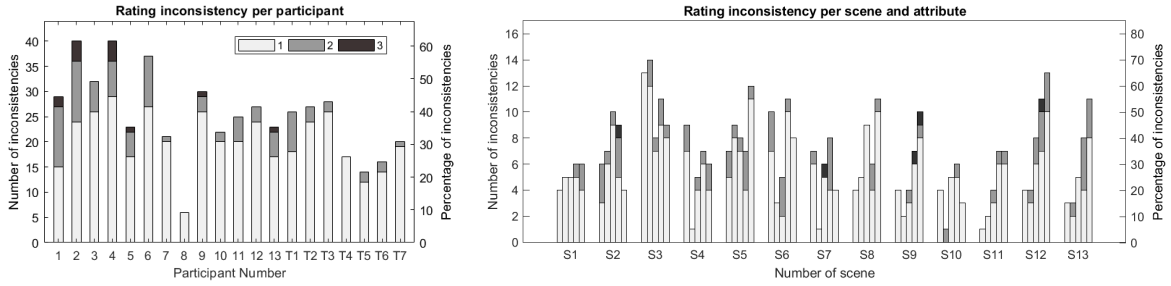


Figure 4.6.: Consistency of the rating in the repetitions for each participant and per scene and attribute

Statistical analysis

This study is based on ordinal scales and samples including repeated measurements. The non-parametric Friedman-Test was used to analyze the results for significant differences between the various test cases. It indicated significant differences ($p < 0.001$) separately for each of the five attributes. The paired, two-sided Wilcoxon signed-rank test served as post-hoc test. The significance level was corrected for multiple test, using the modified false discovery rate (mFDR), because it is less conservative than the Bonferroni correction [243]. Each of the twelve test cases S2-S13 was compared to the reference S1. The corresponding mFDR-corrected significance level is $\alpha_{corr} = 0.0161$. Comparing each scene with each other scene would result in 78 comparisons. The mFDR-corrected significance level would be $\alpha_{corr} = 0.010$. This subsection discusses the observations for each of the five attributes one by one.

Continuity In the previous experiment, no significant differences were found between all scenes with the speech signal. Thus, continuity was not expected to vary significantly in this new experiment. However, it was still included in the questionnaire to keep track of the perceived continuity, on the one hand, for the validation of the technical setup, and on the other hand, to capture potential

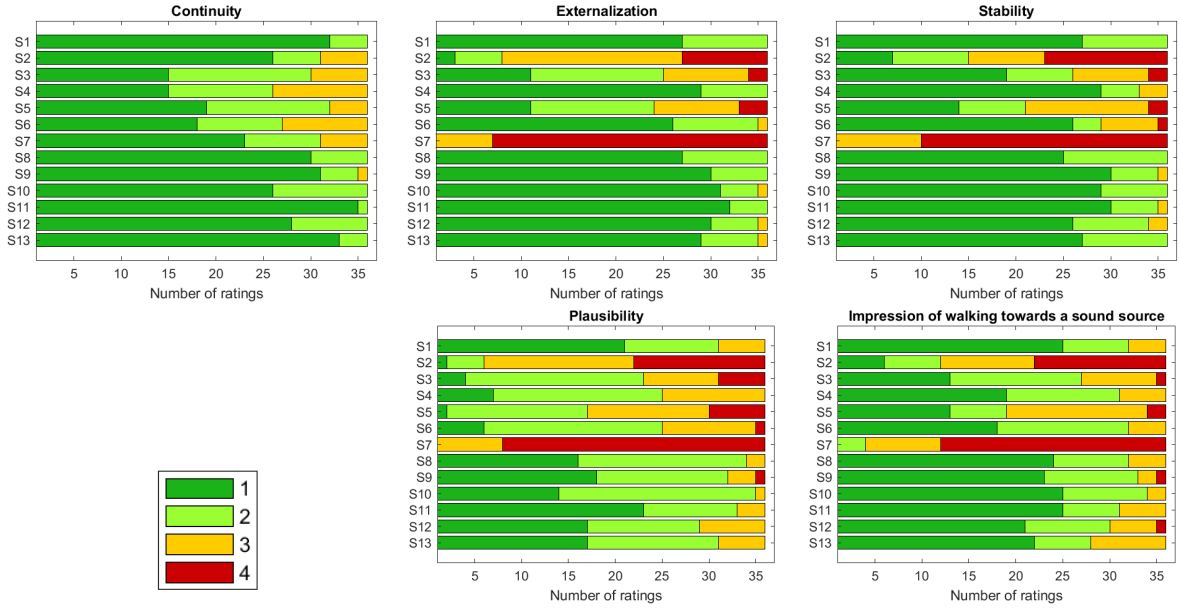


Figure 4.7.: Results for continuity, externalization, sound source stability, plausibility and the impression of walking towards a sound source.

unexpected effects.
The first graph in Fig. 4.7 visualizes the ratings for the different scenes. As expected, the ratings are rather good throughout all scenes. According to the statistical test, the continuity was significantly different for scenes S3 to S7 compared to the original S1 (each $p < 0.009$).
S11 achieved the best ratings for continuity but did not vary significantly from S1 ($p = 0.38$). Adjusting the overall level of the reproduction according to the distance law, as in scenes S3 to S6, seems to affect the perceived continuity. If the level of the full BRIR was varied according to the required adaptation of the direct sound, the overall energy of the BRIR increases considerably from one position to the next. This leads to noticeable steps in reproduction and degrades continuity. Several participants reported noticeable steps in the sound level. Applying the distance law to the reverberation is physically wrong and leads to audible errors. This effect was not observed in the previous study with the BRIRs of the listening laboratory. There, the reverberant energy was very low, and a level adjustment of that reverberation was not severe enough to significantly impact the perceived continuity.
For S7, the continuity was rated significantly different from the original data set. S7 does not contain any reverberation and several subjects reported that there was not enough acoustic change over the different listening positions in both scenes. Some participants degraded continuity due to that, although no discontinuities were noticed.

Externalization The externalization was rated as perfect in most cases. Again, S11 was the scene with the best result and did not vary significantly from S1 with regard to externalization. In contrast, for S2, S3, S5 and S7 externalization was rated significantly worse than for S1 with each $p < 0.001$. For S7, the majority of the subjects chose the answer 'Not external at all.' S7 only contains direct sound but no reverberation. It is well known that reverberant energy is essential for externalization. However, not only the presence of some reverberant energy is essential. Scene S2 was rated as perfectly external in the previous experiment, which was conducted in the listening laboratory. In the present study, most subjects stated that the sound source in the same scene was 'External with larger constraints' or 'Not external at all'. The perception of binaural reproduction is known to be context-dependent. Under the listening conditions given in this experiment, S2 is not in line with

the listeners' expectations. Besides the room in which the test was conducted, the other test scenes were different from the previous experiment since they were based on a different measurement.

Moreover, S3 and S5 were rated as significantly less external. Both scenes were created from the BRIRs measured at the closest position and exhibit the highest DRR. The DRR does not change along the line, but the overall amount of the reverberant energy decreases with decreasing direct sound level.
For S4 and S6, the DRR remains constant, but in those cases, the externalization was not affected. Both scenes are created from the last position in line with the lowest DRR. There, the reverberant energy is stronger than the direct sound energy. With an increase in the overall level, the reverberant energy also increases. In those scenes, the reverberation properties are quite different from the measured BRIRs. Especially in the frontal part of the line, the reverberant energy is considerably stronger than in the real room, but the externalization was not affected.
These observations suggest that more reverberant energy is beneficial for externalization. However, it does not perfectly match the real room, and although the DRR does not change along the line, the reflection pattern remains constant along the line as well. No significant differences were found between S3 and S5 (p=0.69), as well as between S4 and S6 (p=0.22). This suggests that considering the predelay does not make a difference concerning externalization.

Stability of apparent source position The third graph in Fig. 4.7 visualizes the ratings for the stability of the virtual sound source during self-motion. No significant differences could be found between the original data set S1 and scenes S8 to S13, as well as S4 and S6 (each $p > 0.09$).
S7 was significantly different from all other scenes (each $p < 0.001$). Most participants stated that the sound source was moving during their motion. Some reported that the source was moving along while trying to walk away from it. This observation is in line with the findings from the previous experiment. There was a similar test case with only direct sound. The reverberation, especially the progress of the DRR, plays an important role in auditory distance perception, and so it seems to do for the perception of distance change. Even untrained listeners noticed that the auditory distance to the source did not change in accordance with their walking motion.
S2 was significantly different from all other scenes as well (each $p < 0.005$). Again, several participants reported that the source was moving along. This is in line with the results of externalization. If the source is localized in or near the head, the expected progress of the distance change while walking towards it will be distorted. In the previous experiment, source stability was not tested, but S2 was rated as perfectly plausible. This suggests that the source was perceived as mostly stable in the other study. Thus, the context does not only affect externalization but also sound source stability.
Moreover, the sound sources in scenes S3 and S5 were perceived as significantly less stable compared to the original data set (both $p < 0.005$) or the scenes in which only the direct sound was adapted to the changing distance ($p < 0.005$). One trained participant, who had no information on the kind of manipulations, stated to have the impression of a loudspeaker in the constant distance, which was turned louder. Another person had the impression of being in an anechoic room with a loudspeaker, which played back reverberant speech and varied in sound level. As discussed, several people experienced reduced externalization, affecting the perception of the changing distance.
Interestingly, despite the constant DRR throughout the line, the sources in S4 and S6 were perceived as stable as in the original data set. Again, the consideration of the predelay did not cause any audible differences.

Plausibility and the impression of walking towards the sound source As in the previous study, the impression of walking towards the sound source was evaluated separately from the plausibility, although a high correlation between both was expected. The main reason is the challenge of asking inexperienced listeners directly for the plausibility of a virtual sound scene. The term plausibility

was explained to the participants, and the discussion helped to resolve the lack of clarity. Still, an additional question was added to check whether the expected correlation occurs. Fig. 4.7 shows the results.
According to the statistical test, the differences in the impression of walking towards a sound source in scenes S1 and S8 to S13 and S4 and S6 are not significant (all $p > 0.13$). In S2 ($p < 0.001$), S3 ($p < 0.013$), S5($p = 0.002$), and S7 ($p < 0.001$), the impression was rated significantly lower than for the other scenes. Those are also the scenes in which the externalization and the sound source stability were affected. Consequently, the impression of walking toward a sound source is also reduced.

The ratings were significantly lower for scenes S2 to S7 ($p < 0.004$ for each of them) than for the initially measured BRIR set S1. A substantial deviation from the DRR progress in the measured scene, like in S2, S7, or scenes S3 to S6 with constant DRR, seems to affect the plausibility and the impression of walking toward the source. It is interesting to notice that for Externalization, Sound source stability, and the Impression of walking towards a sound source, S4 and S6 were not significantly different from S1, but for continuity and, consequently, for plausibility.

4.2.4. Discussion and Conclusions

Suitability of the method to evaluate plausibility In the previous experiment, plausibility was rated as a measure of agreement with the listener's internal reference. This assumes that listeners understand how walking towards a sound source sounds like, or in this case, towards a loudspeaker and away from it. The previous experiment suggested this test approach due to the challenges and limitations of the alternative test method proposed by Lindau and Weinzierl [211], which were discussed.
Relying on the pure internal reference of the participants also has some drawbacks. Especially inexperienced listeners' internal reference is at risk of being fuzzy or wrong. The current experiment confirmed the related observations from the previous study [81]. For example, several inexperienced subjects did not perceive the initially measured scene as plausible in the room (S1). As a reason, they stated that the change of the loudness is not sufficient in the first half of the translation line and that it works only close to the source. Similar observations were documented for scenes like S8 to S13, which were plausible for most participants. Furthermore, the same people rated scenes like S4 as "clearly plausible" because the loudness change over the given distance was substantial. This suggests that these candidates have inaccurate internal references. With increasing experience, listeners sharpen their internal reference, for example, by paying more attention in the real world or even already during the experiment by listening to the other scenes and starting to appreciate them. Overall, none of the manipulated scenes was rated significantly better than the measured reference scene S1 for any attributes. Consequently, the chosen method with the measured reference seems suitable for most participants.

Role of the selected acoustic manipulations As expected, the BRIR set originally measured in the room in which the experiment took place, later on was rated as plausible by most participants. Furthermore, the results confirm that a realization of the walk-through based on constant reverberation over the tested translation path did not affect plausibility. This holds for all the versions of it, based on the BRIRs measured at either end of the line and also with the original direct sound as well as the simplified realization with pure adjustment of its sound level.
As expected, for creating a plausible approaching motion, the adjustment of the level of the whole BRIR is not sufficient. It is the same as adjusting the level of the sound source while keeping the BRIR unchanged. The DRR does not change in this case, which is an important acoustic cue for the distance change. This was noticed even by many inexperienced listeners.
Interestingly, if the scene were created from the 3.25 m BRIRs, participants were much more tolerant

with regard to keen simplifications, as S4 and S6 were based on a pure level adjustment of the full BRIR. Position 3.25 m is outside the critical distance, and the reverberant energy is stronger than the direct sound energy. This suggests that adding a minimal amount of reverberant energy compared to the real room could support the acceptance of simplified auralization for listener translation.
In summary, pre-delay, ITDG, and the spatiotemporal reflection pattern do not play an essential role in the considered walking path. However, a sufficient amount of reverberant energy and the relative change of the direct sound level, reflected by the DRR, are essential. This is in line with the knowledge about human auditory distance perception summarized in the beginning.

Context-dependencies in the perception of distance changes The BRIR data set measured in the listening lab was rated as plausible by all participants in the previous experiment, which was conducted in that same lab. In the new experiment conducted in a seminar room of similar size but with a considerably longer reverberation time, the Lab-BRIR set (S2) did not satisfy the expectations of most participants.
In both experiments, the same neutral visual scene was reproduced in the head-mounted display, and the same sample of male speech was used as the test stimulus. Thus, the whole audiovisual presentation of the scene was equivalent. Still, plausibility, the impression of walking towards a sound source, the externalization, and even the continuity of this scene were rated substantially worse in this study. In addition, the perceived stability of the sound source during listener motion was not satisfying for S2. In the previous experiment, stability was not rated.
Udesen et al. [246] observed a similar effect for externalization in two different rooms and proposed that the changing visual room impression would be responsible. In the present study, the participants saw the seminar room before the experiment. Although they did not see it during the test, an image of the room might still have been present in their minds.
Werner et al. [247] proposed the theory that the externalization of virtual sound sources is generally degraded by acoustical deviations between the reproduced room and the actual room (listening room). This includes the claim that externalization will also be affected if the reproduced room is "too reverberant" in comparison. This cannot be confirmed in the present study. Scenes S4, S6, S9, S11, and S13 were created from the BRIRs measured at the 3.25 m position, which is located outside the critical distance and, thus, generally more reverberant. All these scenes have not been perceived as significantly degraded compared to the measured reference.
For scenes S7 and S2, many participants reported that the source partly moved along when they tried to walk away from it. That is probably a result of an insufficient change in the auditory distance to match the expected change for the walked distance.
Besides the room in which the experiment was conducted, other conditions changed between the two studies, such as the test scenes. According to the idea of adapting to a room after a certain time, it is likely that the scene tested before the one of interest has an impact. Furthermore, the internal reference may be affected because the selection of test scenes is dominated by scenes that are very similar to the seminar room measurement. A listener's expectations may also be dominated by the seminar room scenes. The lab scene may be perceived as an exception with severe differences to most other scenes in the test, which is therefore somehow wrong.
It would be interesting to repeat the experiment in the seminar room, using the test scenes from the first experiment conducted in the listening laboratory, which are mostly very similar to the BRIR set measured there. Furthermore, constructing a scene from the lab-BRIRs, which has a similar DRR progress to the one measured in the seminar room, could help to assess the role of the DRR in more depth. In general, a better understanding of the process of adapting to the acoustics of a new room is of interest in this context.

Limitations of the study The scene created from a set of BRIRs measured at the given positions in the room serves as the reference which cannot be assumed to be perfectly plausible. It is realized with a step-wise adjustment of the headphone reproduction over the position changes rather than a continuous change. Thus, this scene does not perfectly represent the progress of the natural sound field and the corresponding sound pressure at the listener's ear drums. This limitation seems reasonable since this study focuses on the perceptual effects of the BRIR manipulation and simplification. All of the scenes were represented with the same rendering setup. Therefore, the observed perceptual differences are only due to the BRIR manipulations.

The investigation does not consider authenticity. The manipulations rated as plausible as the reference may still exhibit audible differences in (direct) comparison to the real sound field.
Only a short translation line in front of the source is considered. If the line is located e.g. behind the loudspeaker, the exact progress of the early reflections may be more important because lower direct sound energy increases the audibility of reflections [102].
The considered translation line is also close to the center of the room, and the source orientation is not parallel to any of the walls. Large reflecting surfaces are only present at a certain distance.
Moreover, this study focuses on a listener's approaching motion. The relative angle between the center of the listener and the center of the source does not change. In a passing motion, the directivity of the sound source will play an essential role in the progress of the DRR along the listener's walking trajectory. Such effects are not considered in the presented experiment and remain to be studied in future investigations.

The visual information presented to the subject during the experiment did not provide any source-related or room-related visual cues about the acoustic scene to minimize their potential influence. However, audiovisual plausibility is of interest for applications with a visible object representing the sound source. It demands audiovisual coherence. This is not covered by the experiment. Further studies are necessary.
Only dry male speech was used as the test signal. In the previous study [81], pink noise was used in addition but showed bad results for continuity and, consequently, for plausibility. Consideration of other signals is of interest.

Conclusions

In a distance change of 2 m in front of the loudspeaker, keeping ITDG or spatiotemporal reflection pattern constant did not affect the plausibility of an interactive approaching motion towards the virtual sound source. This holds for both considered rooms and is still the case if the listener crosses the source's critical distance. In both rooms, it had no measurable perceptual impact whether the reflection pattern was taken from the closest or the farthest position. It seems reasonable to assume that any reflection pattern along the line could be used.
The plausibility is not affected by using the same sample of direct sound and adjusting only its level. The adjustment of the pre-delay does not seem to play an essential perceptual role in binaural audio for walking listeners. With faster motion, the pre-delay might become relevant.
In scenes created from BRIRs measured beyond the critical distance, listeners seemed to be less critical with regard to strong simplifications. Even for scenes with constant DRR (S4 and S6), no significant effect on plausibility could be observed.
In contrast, besides the pure lack of reverberation, the reverberation captured in another dryer room also caused significantly lower ratings for externalization, stability, plausibility, and continuity. The experiment contains a scene or BRIR set (S2) rated as plausible in the previous experiment in the listening laboratory, where it was measured. This scene was severely degraded in plausibility in the present study conducted in the more reverberant seminar room. This highlights the impact of context on plausibility.

4.3. Experiment III: Minimum BRIR position grid resolution

This section is based on the conference paper "Minimum BRIR grid resolution for interactive position changes in dynamic binaural synthesis" [326] that I published together with Boris Reif at the 148th AES Convention, Online, in 2020. This work was honored with the Best Student Paper Award of the Convention.

In the past, different approaches for position dynamic binaural reproduction based on interpolation and extrapolation have been proposed. However, which resolution is required to create a smooth reproduction without interpolation is still not known. This information is, for example, of interest to determine the required update rate for position changes when working with purely measured data, which is desirable for studies on perception and, in particular, for evaluating authenticity. Furthermore, if a smooth reproduction can be achieved without interpolation, such scenes are interesting to serve as a reference when evaluating interpolation and extrapolation methods.

Except for the case of walking straight towards a sound source or away from it, the relative angle between the source and listener changes during translation. Thus, the required resolution is likely to be linked to the minimum audible angle (MAA). Perrott and Saberi [248] determined an MAA of 1° in the horizontal plane and 3.6° in the vertical plane. This is in line with values reported by Blauert [1]. Those studies assumed a static listener as well as a static sound source.
The minimum audible angle movement (MAMA) is the smallest detectable azimuthal displacement of a moving sound source relative to a static one [249]. The MAMA increases with velocity and depends on the bandwidth and spectrum of the signal [178]. Results for broadband stimuli suggest that the performance at lower velocities is similar to that obtained with static sources [250]. For angle changes in the horizontal plane, Saberi and Perrott [250] determined a MAMA of 1.7° at a low velocity of 1.8°/s and about 10° at a high velocity of 320°/s. At least for an angular velocity between 25°/s and 100°/s the reverberation does not appear to have an impact.

Brimijoin and Akeroyd [251] observed that in the case of active head rotation, the MAMA is about 1-2° smaller compared to the case where a source moved. However, the relative change between the source and head was identical.
Rummukainen et al. [252] studied the minimum audible angle induced by self-translation (ST-MAA). It was found to be 3.3° in the horizontal plane in front of the listener. In comparison, for a static headphone-based reproduction using interaural time and level differences from a spherical head model, the authors determined an MAA of 1.3°. This suggests that self-translation impairs the absolute localization of stationary sound sources.

With regard to the perception of a continuous change of distance, there are still open questions. Just-noticeable-differences were determined for the DRR [253, 139] as well as the sound level [254, 255]. Those studies only considered static differences. To our knowledge, a minimum audible movement distance (MAMD) has not been investigated yet.

Our investigation intends to provide an evaluation of the perceived continuity in the case of actively moving listeners who explore the virtual sound field by walking and turning around arbitrarily. So far, few investigations consider the exploration behavior. However, the speed of motion will influence the perceived continuity in analogy to the variation of the MAMAs in the case of the moving source. Due to the variation of the MAMAs with the velocity and the spectral properties of the signal, an evaluation of the speed in an intuitive exploration behavior is of interest. Furthermore, noticeable steps may not only be due to exceeding the MAMA or MAMD at that speed but also differences, e.g., in coloration in directly switching to the next filter.
Vorländer [47](p. 100) discusses the perceptibility of spatial variations based on the physical coherence of the room impulse responses of neighbored positions. Under the assumption of an ideal diffuse sound field, the sound pressure p_1 and p_2 at two positions with a distance d can be calculated according to eq. 4.2

$$\Psi_{p_1,p_2} = \frac{\sin(kd)}{kd} \tag{4.2}$$

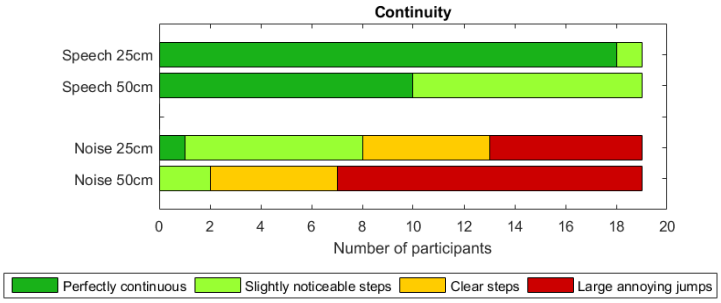


Figure 4.8.: Ratings for the continuity of an interaction approaching motion towards a virtual loudspeaker in a BRIR-based dynamic reproduction with positional resolutions of 25 cm and 50 cm with male speech and white noise (Sec. 4.1).

k is the wave number which depends on the frequency. Based on this equation, Vorländer concludes that for signals with a dominant frequency content between 50 and 500 Hz, a spatial resolution below 10 cm is not necessary. However, also for physical coherence, the just-noticeable differences are not known. Moreover, Vorländer discussed this phenomenon in the context of larger rooms like concert halls and large distances to the sound source. For position dynamic binaural reproduction, scenes in smaller rooms are of interest as well. The perceived continuity will be dominated by the direct sound and early reflections, which cannot be regarded as diffuse. Generally, an ideal diffuse sound field is hardly achieved in real rooms.
The following section summarizes previous studies on the required angular and positional resolution of BRIR filter sets.

Previous studies on BRIR resolution In the case of individually measured BRIRs, with speech, an angular resolution of 2° was sufficient to realize a binaural reproduction, which could not be distinguished from the real auditory scene [29]. However, with white noise reproduced over the same setup with the same configuration, the auralization could be identified by most participants. Lindau et al. [256] observed that an angular resolution of 5° was suitable for music.

The positional resolution has not been investigated in detail yet. Wefers et al. [238] presented a system to create a virtual acoustic environment with binaural technology over loudspeakers using cross-talk-cancellation techniques allowing interactive listener translation. During the position change, the BRIRs are updated section-wise. While the system refreshes the filter for the direct sound after a few centimeters of position change, the reverberant part is updated for changes of 1 m. This approach seems to provide satisfying results in terms of continuity.

The experiment described in Sec. 4.1 studied the perception of an interactive approaching motion towards a virtual loudspeaker realized with a BRIR-based reproduction over headphones with positional resolutions of 25 cm and 50 cm. Figure 4.8 shows the ratings for the continuity of the 19 participants. The test was conducted with male speech and white noise. For white noise, the participants noticed discontinuities in both resolutions, while for speech, a 25 cm-grid seemed to be sufficient. This study focused on a translation motion in which the relative angle between the listening position and the sound source did not change.

The experiment presented in this section investigates different uniformly sampled positional grids for listener translation with regard to perceived continuity. The study considers a listening area also allowing relative angle changes to the source position within a specific range. Furthermore, a variety of source signals are taken into account because of the considerable impact of the signal type in the previous study.

4.3.1. Methodology

Simulations of the BRIR data set

A BRIR data set with a high positional resolution is required for the study. Creating such a data set only by measurements is very time-consuming, and it is hard to place a head-and-torso-simulator (HATS) with the required accuracy. Consequently, the BRIR data set was generated with room acoustic simulations.

The tool MCRoomSim [233] was used to simulate BRIRs as described in Sec. 3.3. The properties of the simulated room were chosen following the properties of the listening laboratory, where the experiments were conducted later. The room has a size of 8.4 m×7.6 m×2.8 m, and the measured BRIR data set is explained in Sec. 3.2. The direct-to-reverberant-energy-ratio (DRR) and T_{30} were extracted for comparison with the simulation. The absorption coefficients of the shoebox model were adjusted until the DRR and T_{30} along the line matched those of the measured data set. Table 4.1 provides an overview of the final values.

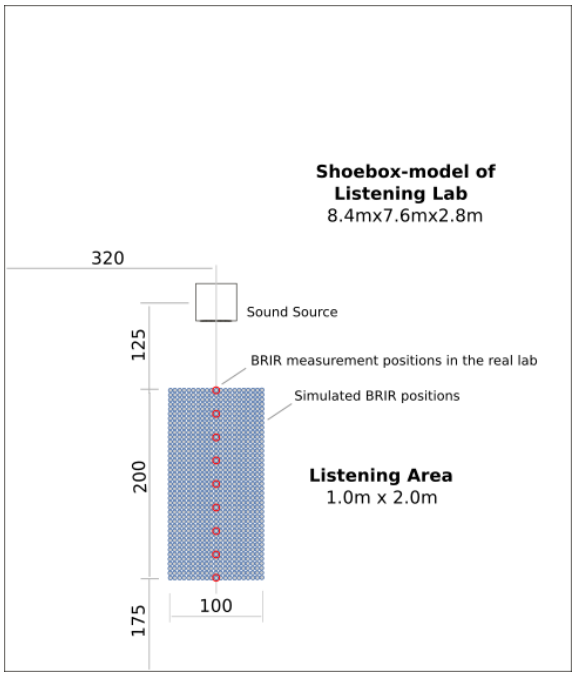


Figure 4.9.: Arrangement of the 1 m × 2 m listening area and the simulated BRIR positions in the virtual room that was constructed as similarly as possible to the original listening laboratory of the university in Ilmenau. The red circles highlight the BRIR measurement positions in the real lab.

The scattering coefficients were set to 0.4 for all walls and frequencies. Similar to the measured listening laboratory, the resulting simulated room had a reverberation time of T_{30} = 0.27 s. A quite dry room was chosen because, with increasing reverberation, the sensitivity to small changes in the relative position of the direct sound might decrease. On the other hand, without any reverberation, the plausibility would be affected, as shown, for example, in experiment 4.1.
The virtual listening position was varied in a positional grid with a resolution of 5 cm over an area with a size of 1 m × 2 m. The arrangement of the source position and the various listening positions in the room is illustrated in Fig. 4.9. At each point, the virtual KU100 dummy head was rotated by 360° in steps of 4°.

With this simulated BRIR data set, two experiments were conducted to evaluate the perceptual consequences of the varying positional resolution. The angular resolution was kept at 4° throughout

	Frequency in Hz					
	125	250	500	1000	2000	4000
wall1	0.5	0.6	0.5	0.4	0.5	0.5
wall2	0.7	0.5	0.4	0.5	0.6	0.4
wall3	0.6	0.5	0.3	0.6	0.6	0.6
wall4	0.5	0.3	0.5	0.6	0.5	0.5
floor	0.5	0.4	0.5	0.3	0.4	0.4
ceiling	0.6	0.5	0.6	0.5	0.5	0.4

Table 4.1.: Absorption coefficients of the walls, floor and ceiling of the virtual room. The scattering coefficients were set to 0.4 for all walls and frequencies.

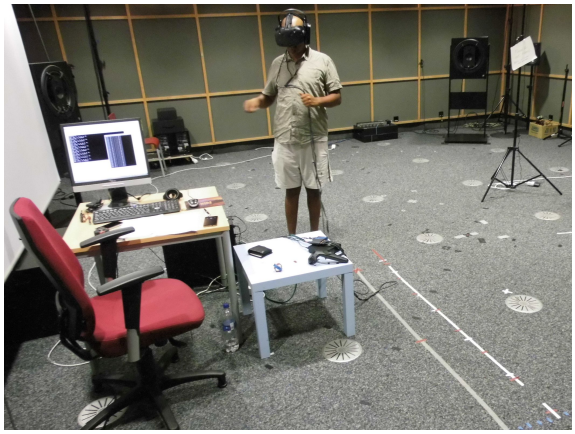


Figure 4.10.: Test setup used in the experiment. The listening area is located around the white line marked on the floor. The participant had to wear an HTC Vive HMD showing a neutral grid-scene. The HMD is tracked in position and orientation. STAX headphones reproduced the audio.

the whole study. In the following sections, both experiments are presented, and the results are discussed in detail.

The reproduction setup

Fig. 4.10 shows the test setup. The participants had to wear the HTC Vive head-mounted display (HMD), which was used for tracking and showed the neutral grid environment explained in Sec. 3.4.1. The software *pyBinSim* (see Sec. 3.4.3 carries out the partitioned convolution of the source signal with the BRIR filter, selected corresponding to the current position and orientation of the listener's head. A blocksize of 256 samples at a sampling frequency of 48 kHz was used for the full length of the filters and the signal. When switching filters, a cosine-square cross-fade (full half-window-overlap) with the duration of one block is applied.
Headphones STAX SR202 driven by an SRM-252II amplifier were used with compensation filters created with the least squares criterion approach described in [53] with roll-off frequencies of 80 Hz to 20 kHz.

4.3.2. Listening experiment - part 1

In the first experiment, the participants were asked to explore the listening area by taking a few steps around. The different grid resolutions (5cm, 10cm, 15cm, 20cm, 25cm, and 100cm) were tested in a randomized order. A looped 8 min excerpt of the audiobook "Sherlock Holmes," read by a dry male voice, was used as a source signal. Fig. 3.7 in Sec. 3.4.3 shows the spectrogram of a 5 s excerpt.

Before the actual test, the highest (5 cm) and the lowest resolution (100 cm) were presented to the participants to give them an idea of the range they could expect in the experiment. Once the test started, all scenes were reproduced in a randomized order. The listener did not know which grid resolution was active. Each resolution was included exactly twice. For each case, the subjects had to rate the continuity on a 5-point scale according to Likert [257]:

- 5 - very good
- 4 - good
- 3 - fair
- 2 - bad
- 1 - very bad

The test method used in this experiment corresponds to the absolute category rating (ACR) - method.

Participants

Twelve non-paid volunteers, four females and eight males, with an average age of 28.4 years, participated in the experiment. Only a few of these participants had previous listening experience. All stated to have normal hearing abilities.

Procedure

The subjects were encouraged to explore the virtual scene by walking around and moving their heads. If they accidentally left the pre-defined listening area for which BRIR filters were simulated, the reproduction was switched to silent mode. The subjects could listen to the scene as long as they wanted. Then, the continuity had to be rated. The experiment took between 30 and 45 minutes per subject. According to reports by the participants, listener fatigue was low until the end due to active exploration.

4.3.3. Results

The exploration duration was very different from person to person and from scene to scene. There was a trend that experienced listeners were faster because they quickly developed an efficient exploration strategy. They could, for example, tell within 3-5 seconds that the 100 cm resolution was not satisfying in terms of continuity.
Each of the 12 participants rated each of the six resolutions twice. That allows for checking the consistency of the participants' ratings. This is in particular important for inexperienced listeners. 72 rating pairs are available for analysis.
Two experienced participants chose precisely the same ratings for the continuity in both trials. Five other participants had an offset of two for one resolution but never for more than once. This offset of two occurred for 15, 20, and 25 cm, never for 5, 10, or 100 cm. There were 17 cases in which a candidate rated a repeated condition with a difference of one. This occurred slightly more often for 25 cm than for the extreme resolution cases.

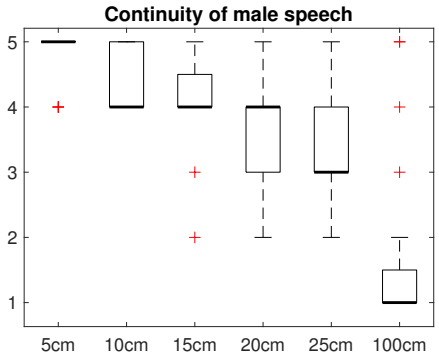


Figure 4.11.: Results of all 12 participants, each resolution was rated twice by every subject for the perceived continuity with male speech.

The results also confirm that inexperienced listeners could rate with relatively high consistency. The few inconsistencies were spread over conditions and participants. Consequently, it was not useful to eliminate any ratings from the analysis.
Fig. 4.11 provides an overview of the ratings of the perceived continuity for the different resolutions. The thick lines mark the medians, and the boxes indicate the 25th and 75th percentiles. The whiskers represent the most extreme ratings except for outliers.

Despite continuous discussions, whether for Likert-scale data, a test for normality is suitable [258], such a test was conducted. According to the Shapiro-Wilk test, the ratings are not part of a normal distribution. A paired, two-sided Wilcoxon signed-rank test was applied to check for significant differences between the conditions. Defining a significance level of $\alpha = 0.05$ for a total of five comparisons (comparing five lower resolutions to the highest) can be achieved by correcting the significance level according to the modified False Discovery Rate (mFDR) to $\alpha_{corr} = 0.219$. If each condition is compared with each condition, mFDR-corrected significance level for 15 comparisons would be $\alpha_{corr} = 0.0151$.

The 100 cm resolution is perceived as significantly less continuous (all $p < 0.0001$) than all other resolutions. This distinct difference is also visible in the boxplot, which is not surprising because, in this case, only six BRIR positions at the edges of the listening area are involved in the reproduction. This leads to a quite obvious switching effect when the listener moves from left to right and crosses the middle line of the area (frontal axis of the virtual loudspeaker). All subjects could notice this except for one. The resolution of 100 cm is insufficient and serves as an anchor condition in the test.
The continuity of the 5 cm resolution was rated significantly better (all $p < 0.001$) than all other resolutions. Consequently, a resolution of 10 cm or lower is not an option, if the best-perceived continuity is desired.
However, it has to be taken into account, that the participants got a short training on detecting discontinuities. They developed an exploration behavior focused on finding such discontinuities by walking very slowly, moving back and forth by a few centimeters at different positions. It is uncertain whether, with a different task, the participants would use the same strategy to explore the scene and whether they would be equally sensitive to small discontinuities.

4.3.4. Listening experiment - part 2

The second experiment aimed to find out which influence of the type of source signal can be expected. For example, from [29], it is known that white noise is more critical with regard to the angular BRIR grid resolution for head rotation than speech. The results of the own experiment

described in Sec. 4.1 show similar observations for the case of translation.
The same BRIR data set, reproduction setup, and test method as in the first experiment were used for this experiment. However, this time three different signals were taken into account: White noise, a solo saxophone, and low pass filtered white noise with a roll-off frequency of 500 Hz and an attenuation of -48 dB per octave. Figure 4.12 shows the spectrograms. This time grid resolutions

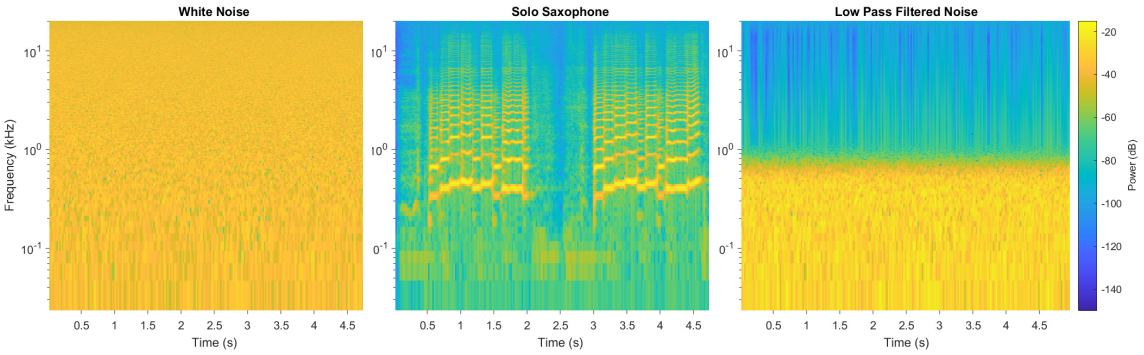


Figure 4.12.: Spectrograms of the three test signals a) white noise - b) solo saxophone - c) low pass filtered white noise.

of 5 cm, 10 cm, 15 cm, 25 cm, 50 cm and 100 cm were chosen for the test.

Participants

25 non-paid volunteers (3 female, 22 male) with an average age of 25.6 years took part in this experiment. 24 reported having no known hearing impairments, one stated having one-sided tinnitus. About half of the participants had previous experiences with psychoacoustic tests in binaural synthesis. Five of the subjects may be considered expert listeners in this field.

Procedure

Before the experiment, white noise was presented with the highest (5 cm) and the lowest resolution (100 cm) to give the subject an idea of the possible range.

In the test, the scenes were divided into three different blocks by the stimulus. The order of these three blocks was shuffled for each participant. Moreover, per block, the six different resolutions were presented in a randomized order. For the perceptual evaluation, the same scale as in the first experiment was used. Each combination of sound source signal and resolution was tested once per participant to keep the duration suitable. In the end, the experiment took between 25-40 minutes per participant. According to the statements of the subjects, they experienced no or very low fatigue effects due to the active motion.

4.3.5. Results

In this experiment, the participants did not repeat any condition. Therefore, the reliability and consistency of each participant's ratings had to be analyzed by other criteria. It could be assumed that a lower resolution will not be better regarding perceived continuity. Thus, the occurrence of better ratings for a lower resolution appeared to be a suitable criterion for the participant's confidence. Six people either showed single extreme opposite ratings or appeared to be judging quite randomly throughout all resolutions. However, the statistical test results remained the same with or without their ratings.
Fig. 4.13 shows the medians with the corresponding 25th and 75th percentiles of the results of the

25 participants. A Shapiro-Wilk test indicated that with a high probability, the data is not part of a normal distribution. The paired, two-sided Wilcoxon signed-rank test was conducted to determine whether there are significant differences between the various resolutions and the best (5 cm). The mFDR-corrected significance level is the same as in Listening experiment part I.

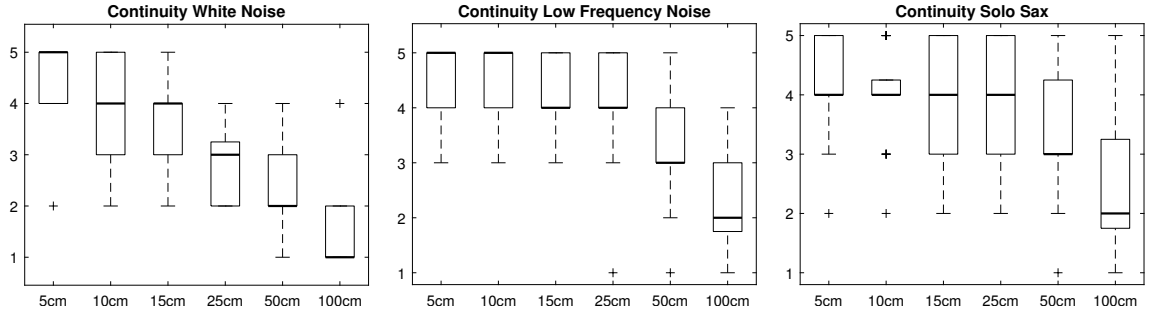


Figure 4.13.: Results of the 25 participants for the perceived continuity with a) white noise - b) low pass filtered white noise - c) solo saxophone.

White Noise It is known that noise is a critical signal with regard to continuity. Most participants rated the continuity for the 5 cm-grid as *very good*. 10 cm is similar with ($p = 0.13$), but for a resolution of 15 cm the continuity was rated significantly lower than for 5 cm ($p = 0.001$). Similar to the ratings given by experienced subjects, the evaluations by naïve listeners improved with increasing resolution. These results suggests that of tested conditions, the 10 cm resolution is the smartest choice for a filter-switching-based reproduction of white noise.

Low Pass Filtered White Noise The low-pass filtered white noise appears to be less critical for lower resolutions. Again, most participants rated the 5 cm-grid as *very good*. However, for this signal, the continuity ratings for the 10 cm ($p = 1.0$) and 15 cm resolution ($p = 0.43$) were not significantly different from that for the highest resolution. For the 25 cm grid, the median is still "good," but most subjects noticed small constraints regarding the continuity (significant difference to 5 cm with $p = 0.011$). With the given realization, 25 cm appears insufficient for the low-pass filtered white noise.

Solo Saxophone The ratings for the continuity for the 5 cm-grid reveal that several participants perceived some constraints. The sax sample started with an intense onset which is critical for the audibility of switching filters. The 18 s long excerpt was looped, so the strong attack was repeated quite often during walking. The ratings for the 10, 15, and 25 cm-grids were not significantly different from 5 cm. Hence, a 5 cm grid is not preferred over a 25 cm resolution for saxophone music. However, even a 5 cm resolution is not ideal.

4.3.6. Discussion and Conclusion

The virtual loudspeaker was placed at a distance of 1.25 m from the listening area. Thus, the occurring maximum angle change induced by listener translation can be calculated as follows.

$$\tan \alpha_{max} = \frac{\Delta d_{grid}}{125\,cm} \tag{4.3}$$

Table 4.2 gives an overview of the corresponding values. Only for the 5 cm-grid, it is approximately in the range of values determined as MAA or MAMA. In most cases, the MAA and MAMA were

determined only for a signal example. The values might be different from the signals used in this test. Furthermore, for practical reasons, an azimuth resolution of 4° instead of 2° was chosen. With the 5 cm-grid, the maximum change of the relative angle between the listener and the source is smaller than the azimuth intervals. It may be argued that this may affect the results. However, for three of the four signals, the continuity was rated as *very good* with the 5 cm resolution. As an evaluation in a practical context, the test setup may be regarded as feasible.

Table 4.2.: Maximum angle change induced by listener translation in the given listening area for the different grid resolutions

d [cm]	**5**	**10**	**15**	**25**	**50**	**100**
α_{max} [°]	2.3	4.6	6.8	11.3	21.8	2*21.8

It could be observed during the experiments that the test subjects developed a particular exploration behavior, searching for audible discontinuities in the reproduction. This behavior is not representative of an everyday listening scenario. The participants moved very slowly. With an increased speed of motion, the sensitivity to discontinuities may be lower. It would be interesting to evaluate the continuity by giving the subjects a task that is not directly related to detecting unsteadiness in reproduction.

Conclusions

The required minimum resolution of the tested uniform BRIR position grid depends on the signal of the virtual source. For male speech and white noise, the highest resolution of 5 cm was rated as significantly more continuous than the 10 cm grid or below. This means that to guarantee the best possible continuity with all signals, a uniform grid with positional intervals of 10 cm or less is not sufficient. For the solo saxophone piece, even a 5 cm resolution did not wholly satisfy the test subjects in terms of perceived continuity, probably due to the transients that are critical with regard to the filter switching procedure, even though a short cross-fade in the time domain was applied to avoid switching artifacts.
Generally, high frequencies are more critical concerning the detectability of filter-switching. The continuous noise, which was limited to frequencies below 500 Hz was the least critical signal in the test. Even for the uniform 15 cm-grid, the perceived continuity was rated as *very good* by most participants. There is probably a certain connection to the spatial sampling, which leads to an upper-frequency limit with regard to aliasing. However, critical continuity was also observed below that frequency. In the case of the 25 cm grid, the upper frequency limit is

$$f_{crit} = \frac{c}{2d} = \frac{343\,\frac{m}{s}}{2 * 0.25\,m} = 686\,Hz. \tag{4.4}$$

The signal is limited to frequencies below this critical frequency, but the perceived continuity is significantly lower than for the 15 cm-grid. The connection is probably more based on the maximal possible phase difference between the two positions. If the source signal is a pure sine with a specific frequency, the perceived continuity will depend significantly on the phase difference between the positions. The most considerable difference could be a switch from the minimum to the maximum or vice versa, which can only occur if there is no sampling point in between. With increasing reverberation, this becomes less and less important. In this experiment, the room is quite dry. Moreover, the effect simplifications in the binaural simulation with the receivers in one position might have on the perception of phase differences between the discrete listening positions is not fully understood. In the end, the results for the low-pass filtered noise in this experiment are not far from the values

determined with the coherence approach discussed by Vorländer [47, p. 100]. For the saxophone, it is more complicated. This approach may provide helpful rough estimations. However, it does not consider effects like the dynamics of the signal or attention and the exploration strategy, for example. After evaluating the continuity, participants explored the virtual environment very slowly to check whether they could detect discontinuities. If the listener's attention is drawn to some other activity or detail, the results may be considerably less critical.

Future work

The simulated room used in this experiment is quite dry, and the listening area was located in front of the virtual loudspeaker. The direct sound was strong compared to the reverberation. Thus, it probably dominates the perception of discontinuities. With stronger reverberation, the reproduction might be perceived as smoother for resolutions rated as critical in this experiment.
This experiment investigated only one positional sampling strategy. As an example, Ajdler et al. [259] discusses different possible approaches for uniform sampling in the context of the mathematical description and physical reconstruction of sound fields. Also, a non-uniform positional sampling might be of interest. Werner et al. [260], for example, suggested the grid getting denser with decreasing distance to the sound source. A psychoacoustic evaluation showed promising results for the tested cases in front of the loudspeaker. However, for prominent strong reflections, this approach might not be satisfying.
Generally, with regard to interpolation and sufficient updating strategies, the required minimum resolution for the reverberant part of the impulse response is of interest. It probably changes with the direct sound level and the energy relation between direct and reverberant sound.

4.4. Experiment IV: Plausibility with and without hidden real reference

This section is based on the journal article "The availability of a real hidden reference affects the plausibility of position-dynamic auditory AR" [275] that I published together with Anna Maria Zerlik in "Frontiers in Virtual Reality" in 2021.

This study was conducted to evaluate our position-dynamic AAR system with the approach proposed by Lindau and Weinzierl [211]. To my knowledge, this was the first time this approach was applied to a system providing interactive walking. Furthermore, this study is of interest to estimate the relevance of including the real version in an evaluation of plausibility. Therefore, an experiment was designed to assess the plausibility of the auditory illusions realized with our AAR system with and without real versions of the scenes among the test items. The following subsection presents the technical realization of the evaluated AAR system, the test scenario chosen for the experiment, and the test design.

4.4.1. Methodology

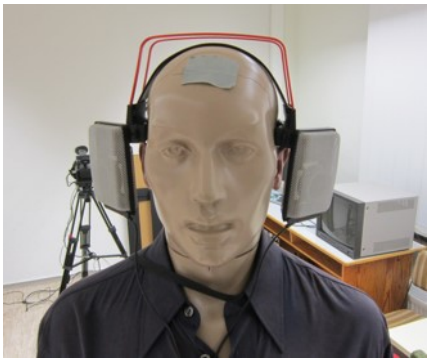
The test scenario was realized in a seminar room of the university in Ilmenau. The participants had to wear headphones. The two loudspeakers in the room could reproduce sound in reality or virtually over headphones. To create the virtual reproduction, BRIR measurements were conducted. The procedure is documented in this section. The test method demands measuring the BRIRs with headphones placed on the dummy's ears to consider their influence on the perception of the real sound field. This influence depends on the type of headphones.

Choice of headphones

Satongar et al. [261] showed that the passive influence of headphones could cause spectral distortions, affect the effective interaural time difference and reduce localization accuracy. Brinkmann et al. [29] used the extra-aural headphones BK211 presented by Erbes et al. [210] for their experiment on authenticity. These headphones may be the best choice for a mixed-reality scenario concerning the lowest impact of the headphone geometry on the perception of the real scene. However, the extra-aural headphones are quite large and heavy. They tend to move slightly on the head during motion despite all efforts to attach them to the listener stably. It may be assumed that these headphones do not allow for a perfectly natural motion. Especially during walking, people may move more carefully to avoid a change of the headphone position on the head. For this reason, we decided not to use extra-aural headphones in this experiment.
Lindau and Weinzierl [211] and Pike et al. [262] used STAX headphones. These cover the ears completely and influence the sound reaching the ears from outside noticeably, for example, by damping the high frequencies. These occlusion or shadowing effects also depend on the direction of the sound incident. In an attempt to find a good compromise, AKG K1000 headphones with an opening angle of 45° on both sides were chosen for this experiment. These headphones are increasingly used for the realization of AR in general. They are less bulky than the extra-aural BK211 and still keep some space between their speakers and the listener's ears. Fig. 4.14a) shows the setup. In the aftermath of this study, we analyzed these effects for different headphones, including all the mentioned ones [278]. Our discussion considers these results.

Measurement of BRIRs

The seminar room chosen for this study has a size of 9.9 m×4.7 m×3.1 m (volume V = 144 m^3) and a reverberation time T_{60} = 0.99 s (broadband). A G.R.A.S. Kemar 45BA with AKG K1000 headphones placed on the ears was set up on an electronic turntable Outline ET 250-3D at nine



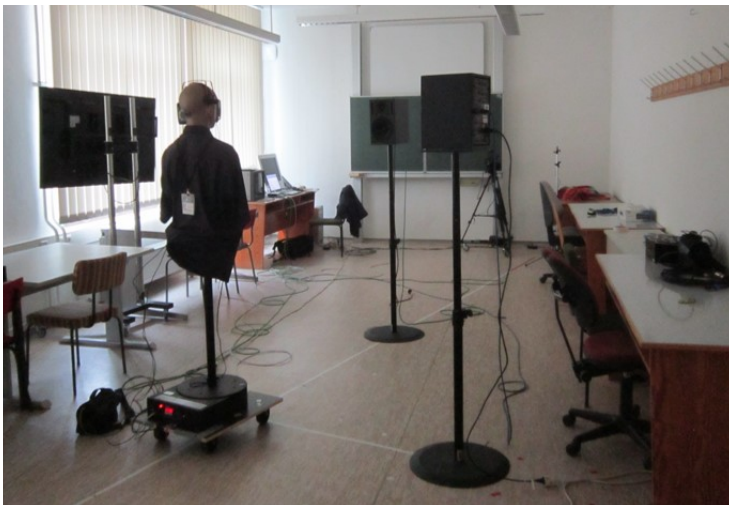


Figure 4.14.: a) AKG K1000 headphones opened by 45° are placed on the Kemar 45BA's ears. - b) Setup for the BRIR measurement in the chosen seminar room.

positions in 25 cm intervals along a line with a length of 2 m. Two loudspeakers Genelec 1030A were positioned in the room, one in front of the line with a distance of 1.25 m to the closest position, one 1.25 m right of the line as illustrated in Fig. 4.15. BRIRs were captured for an azimuth resolution of 2° over the full 360°. Elevation changes were not considered.
We ensured that the headphones did not move on Kemar's head while going through the different positions and head orientations during the measurement. After the BRIR measurement, the headphone transfer function (HpTF) was measured with the exact placement of the AKG K1000. The headphone compensation filter was created from the measured HpTF following the least-squares approaches described by Schärer and Lindau [53]. The captured BRIRs and the created headphone compensation filter are provided as an open-access data set by Neidhardt et al. [319].

Reproduction setup for AAR illusions enduring a position-dynamic listener

After the measurement, the two loudspeakers were kept in the same positions in the same room. An HTC Vive tracker was attached to the headphones to track the position and orientation of the listener's head, as it is shown in Fig. 4.16. The tracking module of the HTC Vive was calibrated to cover the area around the line of measured listening positions. The Python tool *pyBinSim* presented by Neidhardt et al. [317] was used for the partitioned convolution of the dry mono signal with the BRIR filters selected according to the tracking data. The filters had a length of 65536 samples at a sampling frequency of 48 kHz. The block size was set to 512 samples. No interpolation or extrapolation was applied except for a cosine-square cross-fade in the time domain over the duration

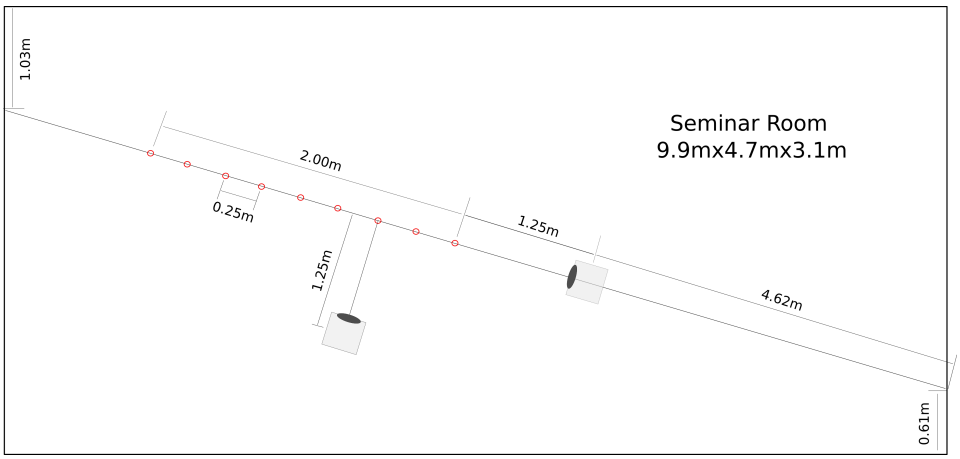


Figure 4.15.: Setup for the BRIR measurement in the chosen seminar room. AKG K1000 headphones were placed on the Kemar 45BA's ears throughout the measurement.

Figure 4.16.: Test person wearing AKG K1000 headphones with a Vive tracker attached to them.

of one block size when switching to another filter. The real-time processing was executed by an Intel Core^{TM} i7-8700K (3.7 GHz) computer with 16 GB RAM and Windows 10 Enterprise (64-Bit). Audio reproduction was realized with an external sound card RME Fireface UCX. The sound level of the two reproduction setups was carefully adjusted by two expert listeners who compared both for several test stimuli.

Individualization of dynamic binaural reproduction

The BRIR filters used for dynamic binaural synthesis contain head-related information like interaural differences in level and time of arrival and spectral characteristics. These physical properties are important acoustic cues in spatial hearing and depend on the individual size and shape of a person's ears, head, and torso. They can vary substantially from person to person. If the binaural reproduction is based on head-related information that does not sufficiently match the listener's head, errors in sound source localization can occur, and externalization can be affected. Both effects may reduce the overall quality of the auditory illusion in terms of plausibility. A wrong match of the individual ear distance can also cause instabilities of the perceived source position during motion. Thus, an individualization of the binaural reproduction is desirable but often requires considerable practical effort like individual BRIR measurements or at least a determination of individual interaural time difference (ITD) combined with an adequate BRIR adjustment. Brinkmann et al. [29] measured individual BRIRs for each participant before evaluating the authenticity of the binaural reproduction. Lindau and Weinzierl [211] conducted their study with two systems based on non-individual BRIRs

measured with a FABIAN dummy head. In one of them, the ITDs were extracted and individually adjusted for each listener. With this system, a plausible reproduction according to the given test paradigm was achieved. For the other system, coloration and unstable localization were reported Lindau et al. [214]. Pike et al. [262] tested the plausibility of dynamic binaural synthesis for head rotation with non-individual BRIRs of a small room with the method suggested by Lindau and Weinzierl [211]. The BRIRs were measured with a Neumann KU100 dummy, but an individualization of the ITDs was realized in the post-processing. Before the test, participants had to determine their ITD by listening to reproductions with different ITDs, which is not an easy task even for experts. With their setup, still slight instabilities in source localization were reported and, e.g., described as increased localization blur or increased apparent source width. In their experiment, a sensory distance between the real sound field and auralization was found.

Participants

Seventeen people aged between 18 and 33 years volunteered to participate in the experiment. The average age was 25 ± 2.57 years. Five subjects were experienced listeners in BRIR-based binaural synthesis, and the others were mostly inexperienced. Experienced listeners were expected to be more critical of plausibility. For this reason, we were interested in recruiting at least a suitable number of them to allow for a separate analysis of this group. All participants were master's students or Ph.D. candidates at the university in Ilmenau and were interested in the field of AR. The selected group is considered representative of users of AR systems. The panel consisted of four female and 13 male listeners. All volunteers stated to have normal hearing abilities without any impairments. All participants completed the whole experiment, and all their results were included in the statistical analysis.

Test scenes

The two different loudspeaker positions were considered as different test cases. Three test signals were included in the experiment:

- Speech: Dry female speech reading an audiobook
- Music: Pop song (Left channel as mono)
- Snare Drum: 50 bpm

Although the loudness of loudspeaker reproduction and binaural auralization was adjusted carefully, two different sound levels (0 dB and -6 dB) were included in the test to minimize the potential influence of minimal loudness differences in the determination process. This adds up to a total of 12 test scenes for each of the two reproduction methods. Tab. 4.3 provides an overview.
All stimuli were band limited to a frequency range between 150 Hz and 16 kHz to reduce the influence of low-frequency background noise and loudspeaker distortion in the high frequencies.

Pre-test with few experts

In preparing the official experiment, a few expert listeners conducted an informal pre-listening session. Both direct AB-comparison and blind identification of auralization and real sound field were part of this procedure. The results and observations are documented in Sec. 4.4.2. During this critical listening session, the experts observed that a fade-in is required after activating the headphone reproduction. An abrupt start of the signal in the headphone reproduction revealed the virtual scene. This was considered in the final experiment.

Source position	Audio content	Gain
Frontal	Speech	0 dB
Frontal	Speech	-6 dB
Frontal	Music	0 dB
Frontal	Music	-6 dB
Frontal	Snare Drum	0 dB
Frontal	Snare Drum	-6 dB
Side	Speech	0 dB
Side	Speech	-6 dB
Side	Music	0 dB
Side	Music	-6 dB
Side	Snare Drum	0 dB
Side	Snare Drum	-6 dB

Table 4.3.: Two different source positions, three types of signals and two different sound levels were taken into account.

Listening experiment with the test panel

Before participating in the study, informed consent was obtained from all individual participants involved in the study. The participant had to wear the AKG K1000 headphones with the Vive tracker in the experiment. At the beginning of each trial, the subject had to stand at the end of the translation line (measurement position with a distance of 3.25 m to the loudspeaker in the front). The participant was told that the real loudspeaker or its binaural simulation would be randomly presented. The task was to decide which of the two versions was currently active. In addition, the subject was instructed to move along the line and use head rotation and self-rotation arbitrarily.
The first part of the test aimed to investigate plausibility with respect to a pure internal reference. For this part, it had to be avoided that the participant gets an impression of the real version of the sound field. Therefore a training session was not feasible. In the second part, real scenes were included as test items to evaluate plausibility with regard to the internal reference tuned by the real versions of the scenes.

Test part I All test scenes in their binaural version, 12 in total, were presented in a randomized order. The real reproduction was not included in this test. This part took about 15-20 minutes per participant.

Test part II All test scenes in their binaural and their loudspeaker version, 24 in total, were presented in a randomized order. This part took about 30-40 minutes per participant.

The participants were asked to evaluate 36 test scenes wherein the amount of virtual and real scenes is not necessarily similar. After the experiment, the participants were asked to describe the audible cues they used to distinguish simulation and real reproduction.

Required sample size and test duration

To achieve statistically meaningful results, an appropriate sample size is required. Furthermore, it is crucial to consider that taking time to explore the scene and make a decision may affect the rate of correct answers. Lindau and Weinzierl [211] conducted their experiment with 11 experienced listeners. Each of them had to evaluate 100 test samples. This allowed for an analysis of the individual sensitivity d'_i, hence, the discriminability, based on the Signal Detection Theory. However, in their experiment, each test stimulus was only six seconds, which was possible because interactive self-motion was limited to $\pm 80°$ in azimuth. The test was restricted to a one-time listening per

sample. The authors reported that none of the participants took longer than 15 min for the whole test.

In our experiment, each of the 17 participants completed 36 evaluations. The participants were allowed to listen to and explore the scene as long as they thought it was helpful. On average, the assessment took the subjects 70 seconds per test scene. Between the scenes, there was a break of 20-25 seconds for the test conductor to take notes and start the new scene. The experiment with introduction and interview at the end took between 50 and 70 minutes. Due to the breaks, the active exploration of the scene, and the reportedly interesting task, listener fatigue was acceptable.

Especially in systems with a high degree of interactivity, there will always be a trade-off between large sample sizes and providing the participants with a suitable amount of time to explore the scene and make their decisions.

Methods for statistical analysis

A standard method to analyze the results of an experiment conducted in a Yes/No-paradigm is based on the Signal Detection Theory. The following paragraph explains how the SDT can be used to estimate the discriminability between real and virtual reproduction.

a) Estimating the discriminability based on Signal Detection Theory The participants have two answering options 'virtual' and 'real'. The type of reproduction can also be both virtual or real. If the participant cannot detect a cue indicating that the virtual sound source is active, the participant is more likely to pick the answer 'real'. Based on this idea, the real sound source is regarded as 'Noise' and the virtual sound source with potential, revealing cues as 'Signal'. In accordance to [263], the four possible outcomes in this classic SDT experiment are called Hit, Miss, False Alarm, and Correct Rejection. Tab. 4.4 provides an overview.

Response	Virtual (Signal)	Real (Noise)
'Virtual'	Hit	False Alarm (FA)
'Real'	Miss	Correct Rejection (CR)

Table 4.4.: Possible outcomes in the Signal Detection paradigm.

The primary goal of SDT is to determine the sensitivity index d' and the decision criterion c. In this specific case, d' is the sensitivity to cues revealing the virtual reproduction as virtual. Thus, a sensitivity $d' = 0$ indicates that the virtual sound source cannot be distinguished from the real sound source. In this case, 'Perfect Plausibility' would be achieved. The sensitivity is a measure of the discriminability of the virtual sound source from the real one. The decision criterion indicates whether there are any tendencies towards one of the two answers.

Using SDT, the most consistent analysis is possible if one observer completes many assignments for the same stimulus in its virtual and real versions. Suppose more subjects and stimuli are taken into account. In that case, the theory demands to determine the individual sensitivity d' for each combination of subject and stimulus separately and then calculate the *mean sensitivity*. If the sample size for each combination is too small, the sensitivity has to be determined for a pool of observers and stimuli. This *pooled sensitivity* is discussed in detail by Macmillan and Creelman [264] [p. 331 ff].

Several previous plausibility studies used the SDT for their analysis, for example, by Lindau and Weinzierl [211]. They calculated the individual sensitivity per person, still averaging over different signals and source positions, then calculated the mean sensitivity. Only the overall percentage of correct answers was taken into account, assuming that the number of correct answers would be equally

distributed over real and virtual scenes and considering equations developed for a 2AFC-test design. A Yes/No paradigm differs from a 2AFC paradigm. In a Yes/No paradigm, the stimuli are presented and rated one by one. In contrast, the 2AFC-paradigm, as it is considered in the SDT, offers Noise- and Signal-Stimuli (in our experiment, real and virtual) within one trial, either in randomized temporal or spatial order. A 2AFC paradigm, therefore, allows for a direct comparison between both stimuli. Furthermore, the answer in each trial is correct or wrong for both stimuli simultaneously. In a Yes/No paradigm, distinguishing between Hits and Correct Rejections can provide additional or more accurate information since they are not necessarily equal. Fig. 4.17 visualizes the percentage of correct answers of our experiment separated by real and virtual reproduction and shows that they are not equal. Therefore, we considered p_{Hit} and p_{FA} rather than only the percentage of correct answers. According to e.g. [265] p_{Hit} and p_{FA} can be calculated as follows

$$p_{Hit} = \frac{\text{Number of Hits}}{\text{Total Number of Signal Presentations}} \qquad p_{FA} = \frac{\text{Number of False Alarms}}{\text{Total Number of Noise Presentations}}$$

The sensitivity d' can be determined with the following equation:

$$d' = z(p_{Hit}) - z(p_{FA}) \tag{4.5}$$

This equation is a criterion-free estimation of the sensitivity. It can be used to determine the individual and the pooled sensitivity. For extreme values of p_{Hit} and p_{FA}, a correction according to Hautus [266] was applied. This correction is integrated into the *dprime*-function in R, which we used for this analysis. Since the sample size per person was relatively small, both mean and pooled sensitivity will be estimated and compared. In addition, the decision criterion location c can be calculated. c indicates the distance of the decision criterion from the center between both distributions.

$$c = -\frac{1}{2}(z(p_{Hit}) + z(p_{FA})) \tag{4.6}$$

c is zero if False Alarms and Misses occur with an equal percentage of the Noise- and Signal samples. If c is below zero, there is a tendency towards the answer *virtual*. In contrast, a positive value indicates a tendency towards the answer *real*.
Another question is, which value of d' indicates that the discriminability of the virtual reproduction is sufficiently small. Lindau and Weinzierl [211] determined such a minimum effect hypothesis under the assumption of non-biased participants and only considering the percentage of correct answers. The determination becomes more challenging for a group of subjects with considerable differences in individual bias. Therefore we additionally consider another interpretation of the data.

b) Analysis based on the paired t-test It is interesting to analyze the rate of acceptance as *real*. This number is equivalent to the number of correct answers for the real source. For virtual reproduction, it is the number of wrong answers. The auditory illusion can be considered plausible if the acceptance rate for the virtual source does not vary significantly from that of the real sound source. In order to test for significant differences in the rates of acceptance between the real and the virtual test scenes, a paired t-test can be used. The t-test is suitable even for small sample sizes. The analysis considers the distribution of the individual rates of acceptance for both test conditions. The paired t-test assumes that the difference between both test conditions follows a normal distribution. This was tested and confirmed with a Shapiro-Wilk test, although it has to be noticed that test for normal distribution can be inaccurate for small samples. The paired t-test checks whether the hypothesis that the two samples follow distributions with equal means can be rejected.

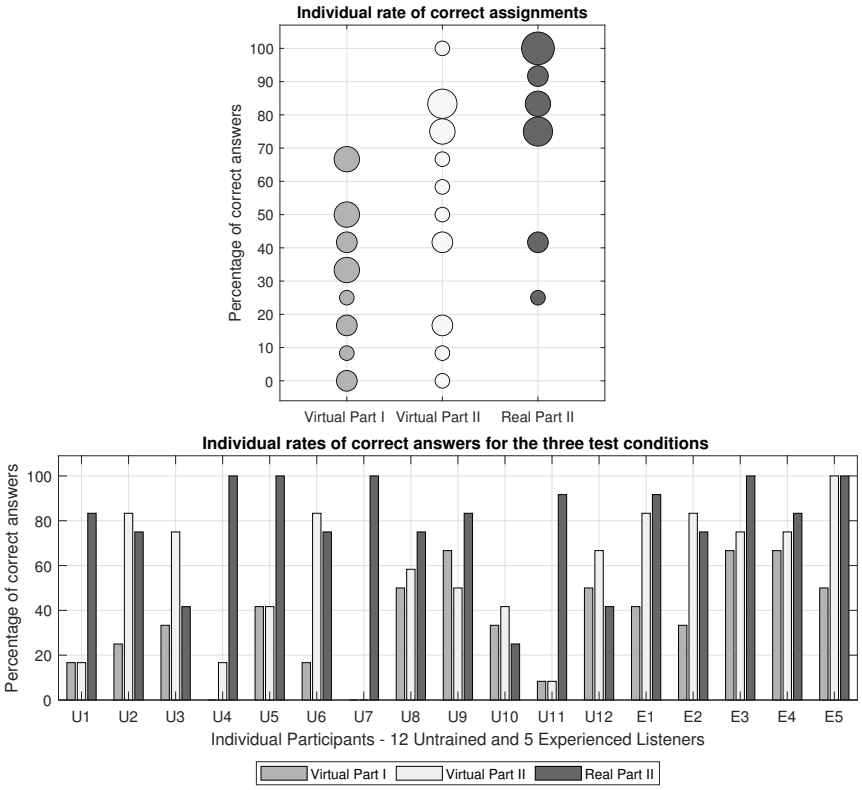


Figure 4.17.: a) Individual rates of correct answers sorted by test condition. The size of the bubbles indicates how many subjects achieved this result. - b) Percentage of correct answers achieved by the 17 individual participants for the scenes with the virtual reproduction in Part I and Part II of the experiment as well as the real source in Part II.

4.4.2. Results

The auditory illusion of a loudspeaker reproducing sound is considered plausible if the listeners cannot identify it as virtual systematically. The realization of the position-dynamic binaural synthesis in this experiment does not contain any individualization of the BRIRs. Consequently, we expected that at least the experienced listeners would detect the virtual reproduction among the test scenes in this Yes/No paradigm. The study also aims at identifying available audible cues that can reveal the simulation. This is of interest for a targeted improvement of the system.
Furthermore, since considering a real reference in a perceptual evaluation comes with practical challenges and limitations, we want to know whether the availability of a real reference influences the estimated plausibility of the auditory illusion. For this reason, the experiment was conducted in two parts. The first evaluates the plausibility regarding the pure internal reference without considering real sound fields. The second part evaluates plausibility with the test approach proposed by Lindau and Weinzierl [211] by including real versions of the simulated sound fields. Does the availability of the real sound field affect the plausibility?

Observations of the informal pretest

In the pre-test, three experts who did not participate in the subsequent main experiment listened to the real and the virtual version of the loudspeaker reproduction in a direct AB comparison for the various test cases listed in Tab. 4.3. The experts described freely which differences they perceived. It was interesting to notice that after a short exploration episode, the experts moved to the closest

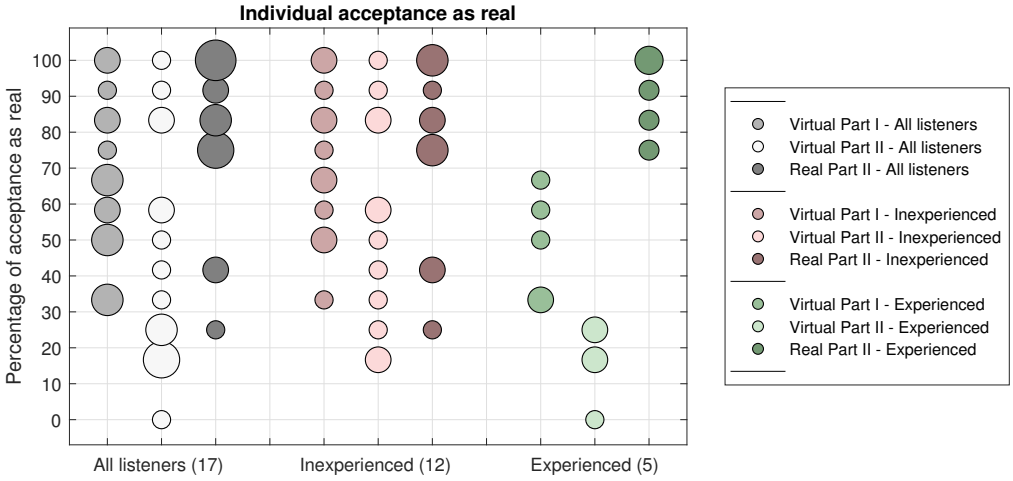


Figure 4.18.: Percentage of test scenes which were rated as *real* by the 17 individual participants, as well as Inexperienced and Experienced Listeners separately for the scenes with the virtual reproduction in Part I and Part II of the experiment and the real scenes in Part II.

position possible to the front of the active sound source. Once they arrived, they focused on rotating their heads or turning themselves in that position. Sometimes, they reported a slight instability of the perceived location of the sound source during head rotation. Furthermore, when turning the back toward the sound source, real and virtual reproduction differences were audible. The deviations were described as a change in distance perception, externalization, and relative sound level. For binaural reproduction, the source was described as in the head or sticking to the back of the head. However, with the real reproduction, the source in the back did not appear entirely natural. The distance perception did also not match the expectations. In the AB comparison, the experts noticed minimal deviations in timbre, reverberance, and apparent source width in addition to the previously mentioned effects.

Overview and individual differences

In this experiment, each of the 17 subjects rated 36 test scenes. In total, these are 612 answers. 348 of these answers (56.9 %) were correct. With 30 correct assignments out of 36 (83 %), one of the trained listeners achieved the highest individual number of correct answers. The other experienced participants rated 23, 27, 28, and 29 scenes correctly in the course of the experiment. Two inexperienced listeners achieved the lowest individual rate of correct responses with 12 out of 36 (33 %). These numbers indicate that the task of identifying the virtual reproduction among the randomized test items was not an easy task. However, the numbers sum up different test cases that should be considered separately. The three main categories of test cases are 'virtual sound source tested in part I of the experiment,' 'virtual sound source tested in part II of the experiment,' and 'real sound source tested in part II of the experiment.' Each of the 17 participants rated 12 test scenes for each of these categories and achieved an individual number x_i of correct answers between 0 and 12.
Fig. 4.17b) illustrates the individual rates of correct answers each participant achieved in the three test conditions. The percentage of correct answers varies substantially among the participants. Furthermore, the distribution of the correct responses over the three test conditions is very different from person to person.

Fig. 4.17a) basically shows the same numbers but sorted by test condition. The data for the separate conditions exhibit different trends. A paired t-test was conducted to test whether the sample of individually achieved rates of correct answers is part of distributions with equal means.

According to the paired t-test, in part II of the experiment for the cases when the visible loudspeaker was actually reproducing the sound, the participants answered significantly, $t(16) = 2.24$, $p < 0.04$) more often correctly ($M = 9.47$, $SD = 2.74$) than for the test scenes with the virtual reproduction ($M = 6.76$, $SD = 3.68$). Furthermore, for the virtual scenes in part II of the experiment, the subjects answered significantly more often correctly, $t(16) = 3.50$, $p = .003$ ($M = 6.76$, $SD = 3.68$) than for the same test scenes in part I ($M = 4.24$, $SD = 2.63$).

Correct identification of the real source and its limitations

First, it is of interest how often the participants identified the real sound source as real. Each of the 17 participants evaluated 12 test cases in which the sound source was real. This results in a total of 204 evaluations. Overall, in only 161 of the 204 assignments (78.9 %), the participant chose *real* as the answer. This indicates that the internal reference is not perfectly reliable for some of the listeners. Probably, most participants have never paid attention to what it sounds like to walk towards or past a loudspeaker or turn around in front of it. Usually, listeners have a basic idea of what to expect but feel uncertain about the details. Additionally, the subjects had to listen to the real loudspeaker while wearing headphones. This is an uncommon listening situation for which most listeners might not have an adequate internal reference. Generally, real listening scenarios may exhibit details that the listener did not expect. Such elements may be mistaken as cues revealing the virtual sound source. Especially for listeners with no or little experience in the field of binaural technology, the task was challenging. The five experienced listeners correctly identified the real source in 12, 12, 11, 10, and 9 of the 12 test cases (on average 90.0 %). Inexperienced listeners were correct in 74.3 % of the cases. Fig. 4.17 visualizes the individual results. Three inexperienced listeners rated the real loudspeaker reproduction as real only in 3 or 5 of the 12 test cases. Especially the person with the three correct identifications tended to assign virtual and real scenes and vice versa.

Analysis of Part II: Plausibility evaluation with a tuned internal reference

This part of the analysis focuses on part II of the experiment, where the plausibility was evaluated, including the real counterparts of the test scenes. This test design follows the method proposed by Lindau and Weinzierl [211]. They determined the sensitivity index d' as an indicator of the discriminability between real and virtual versions of the scenes based on the Signal Detection Theory (SDT). We analyzed our results accordingly.

a) Estimating the discriminability based on Signal Detection Theory The sensitivity index d' can be calculated with eq. 4.4.1. Due to the small sample size per person small, in addition to the standard mean sensitivity, we determined the pooled sensitivity to compare both. The first column in Tab. 4.5 shows the results for part II of the experiment. The mean sensitivity determined from the individual sensitivities of each participant differs only slightly from the pooled sensitivity, which was determined from the overall number of Hits and False Alarms. Both values are close to one and indicate good discriminability. The decision criterion c is determined with eq. 4.6. Due to the small sample size, c was also calculated as a mean of the individual response bias and as the pooled criterion overall. The difference between both values is minimal. The positive value shows that the location of the decision criterion is shifted towards the distribution of Hits. This indicates that in part II, the subjects had, on average, a tendency towards the response *real*.

b) Analysis based on the paired t-test Fig. 4.18 shows how often the participants picked the answer *real* in each of the conditions. This indicates the rate of acceptance as *real*. The paired t-test checks for the hypothesis that the two samples follow distributions with equal means. For the distributions of the individual acceptance rates as real, this hypothesis can be rejected, $t(16) = 4.18$,

Results for d' and c	Part II (tuned internal reference)	Part I (pure internal reference)
mean sensitivity	1.05 ($c = 0.37$)	0.46 ($c = 0.67$)
pooled sensitivity	0.96 ($c = 0.32$)	0.43 ($c = 0.59$)

Table 4.5.: Estimated sensitivity d' and decision criterion c for both parts of the experiment. As expected the sensitivity estimated for part I is considerably lower than for part II. For both parts the decision criterion indicates a tendency towards the response *real*. In part I this tendency is even stronger than in part II.

$p < 0.001$. This means, in part II, the acceptance of the virtual reproduction ($M = 5.23$, $SD = 3.68$) was significantly lower than for the real reproduction ($M = 9.47$, $SD = 2.74$).
In addition to the results of the participants, Fig. 4.18 shows different results for experienced and inexperienced listeners. For both groups, the paired t-test indicates significant differences between the acceptance of real and virtual reproduction, experienced $t(4) = 9.6$, $p < 0.001$ ($M_{real} = 10.80$, $SD_{real} = 1.30$ and $M_{virtII} = 2.0$, $SD_{virtII} = 1.23$) and inexperienced listeners $t(11) = 2.50$, $p < 0.05$ ($M_{real} = 8.92$, $SD_{real} = 3.03$ and $M_{virtII} = 6.58$, $SD_{virtII} = 3.53$). Although the numbers indicate that discriminability is quite good, the subjects found it hard to distinguish whether the loudspeaker was reproducing sound virtually or for real. They had the chance to take as much time as they needed to explore the scene and decide. An average duration of the exploration per scene of 70 seconds indicates that the decision was not taken right away. Providing a convincing auditory illusion of the given scenario that endures this high degree of interactivity and this long and intense exploration is a more critical test than a short one-time listening. Achieving plausibility with regard to a "tuned" internal reference is more challenging.

Analysis of Part I - Plausibility with regard to the pure internal reference

In Fig. 4.17b) the first and second row of bubbles shows the individual percentage of correct identifications of the virtual sound source in the first and the second part of the experiment. In part I, in the case in which only the virtual sound source was presented, in 71 of the 204 test scenes (34.8 %), the virtual sound source was identified correctly. In part II, the virtual sound source was presented alongside the real version in a randomized order. This case was correctly identified in 115 of the 204 scene assignments (56.4 %). The statistical analysis is again based on two approaches, Signal Detection Theory and the paired t-test.

a) Analysis based on Signal Detection Theory In order to compare the evaluations of the virtual sound sources in part I and part II of the experiment in SDT, the sensitivities were calculated for both parts in relation to the evaluation of the real sound source conducted in part II. Thus, mean and pooled sensitivity were calculated again, this time with $pHit$ based on the rate of correct identifications of the virtual reproduction in part I instead of part II. Tab. 4.5 provides an overview of the estimated sensitivities.
Again, mean and pooled sensitivity is very similar. As expected, the sensitivity estimated for part I is considerably lower than for part II. For both parts of the experiment, the decision criterion indicates a tendency towards the response *real*. In part I, this tendency is even stronger than in part II.

b) Analysis based on the paired t-test Considering the individual rates of acceptance as real, it is the question of whether there is a significant difference between the acceptance of the virtual scenes in part I and part II of the experiment. The results of the paired t-test indicate that over all subjects, the hypothesis of equal means can be rejected, $t(16) = 3.50$, $p < 0.005$. In part I the acceptance of the virtual reproduction ($M = 7.76$, $SD = 2.63$) was significantly higher

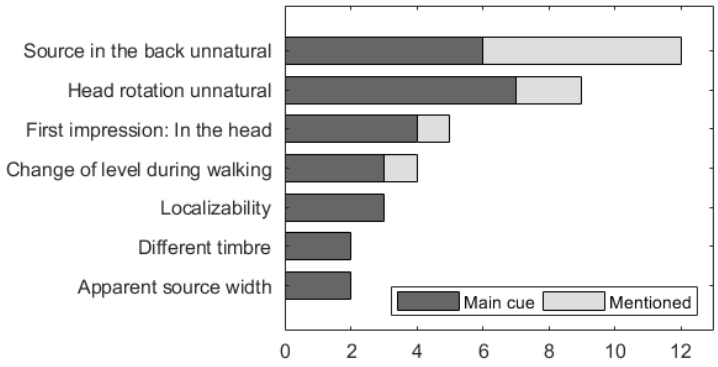


Figure 4.19.: Overview of audible cues reported to be used by the participants to discriminate the binaural simulation from the real sound field.

than in part II ($M = 5.24$, $SD = 3.68$). This holds for both, experienced ($t(4) = 3.28$, $p < 0.04$), ($M_{virtI} = 5.80$, $SD_{virtI} = 1.79$ and $M_{virtII} = 2.0$, $SD_{virtII} = 1.22$) and inexperienced ($t(11) = 2.25$, $p < 0.05$) ($M_{virtI} = 8.58$, $SD_{virtI} = 2.54$ and $M_{virtII} = 6.58$, $SD_{virtII} = 3.53$) listeners. This result is not surprising. An influence of real scenes among the test items was expected.
In addition, it is interesting to compare the results of part I to those of the real scenes. For a significance level $\alpha = 0.05$, the hypothesis of equal means cannot be rejected, $t(16) = 1.94$, $p = 0.07$. Thus, the acceptance of the virtual reproduction in part I of the experiment ($M = 7.76$, $SD = 2.63$) is not significantly different from the acceptance of the real scenes in part II ($M = 9.47$, $SD = 2.74$). This is an exciting observation. Taking only the experienced listeners into account, the paired t-test indicates that the means of the acceptance of virtual scenes in part I ($M = 5.80$, $SD = 1.79$) and real scenes in part II ($M = 10.80$, $SD = 1.30$) differ significantly, $t(4) = 4.23$, $p = 0.01$. The bubble-chart in Fig. 4.18 visualizes the individual acceptance rates for experienced listeners. The rates are visually quite well separated for the three test conditions. For inexperienced listeners, the paired t-test does not reject the hypothesis of equal means at all, $t(11) = 0.34$, $p > 0.7$ ($M_{virtI} = 8.58$, $SD_{virtI} = 2.54$ and $M_{real} = 6.58$, $SD_{real} = 3.53$). This means the created spatial auditory illusion is convincing enough that inexperienced listeners do not notice it is an illusion when relying purely on their internal references. This observation is essential for future studies with the goal of evaluating plausibility.

Cues used for detection of the virtual reproduction.

Fig. 4.19 provides a summary of the audible cues mentioned by the participants in the interview after the test. This overview does not consider the relation to the individual detection rates but represents all answers given by the subjects.
Twelve of the 17 subjects reported that the sound source behaved unnaturally when they turned their backs toward it. The source appeared closer, sometimes even in the head, and varied in loudness. This observation is in line with the reports by the trained listeners in the pre-listening session.
Nine participants reported an unnatural experience of head rotation. The source position appeared slightly unstable. The effect increased with the speed of rotation. Seven of the subjects stated that this was the main cue they used to identify the binaural auralization. This observation is also in line with the effects reported by the experts in the pre-test.
In addition to these two major cues, some participants reported that they perceived the sound source in the head before they started to move. Some listeners mentioned that the sound level changed in a way they did not expect. Few people stated that they perceived differences in timbre, apparent source width, and localizability.

Source position, type of signal, and sound level

Fig. 4.20 provides an overview of the rates of correct answers with respect to the source position, the type of signal, and the sound level. Only part II of the experiment is considered for this analysis. The first graph visualizes the subjects' individual rates of correct answers within the test. For each source position and each sound level condition, the total amount of test cases per person was twelve, six real and six virtual. According to the paired t-test, the number of correct answers for the source in the front ($M = 7.94$, $SD = 2.38$) does not differ significantly, $t(16) = 1.0$, $p > .3$, from the percentage of correct answers for the lateral source position ($M = 8.29$, $SD = 2.02$). The percentage of correct answers for the 0 dB sound level ($M = 8.24$, $SD = 2.22$) was not significantly different, $t(16) = 0.57$, $p > .3$, from that for the -6 dB ($M = 8.0$, $SD = 2.29$).
For each type of signal, each participant rated eight scenes in part II, four virtual and four real. The individual rates of correct answers for speech ($M = 5.29$, $SD = 1.61$), music ($M = 5.47$, $SD = 1.70$) or snare ($M = 5.47$, $SD = 1.94$) were not significantly different from each other, $t(16) < 0.5$, $p > .6$ for all three combinations. In summary, the position of the sound source, the type of signal, and the sound level did not significantly influence the percentage of correct answers.

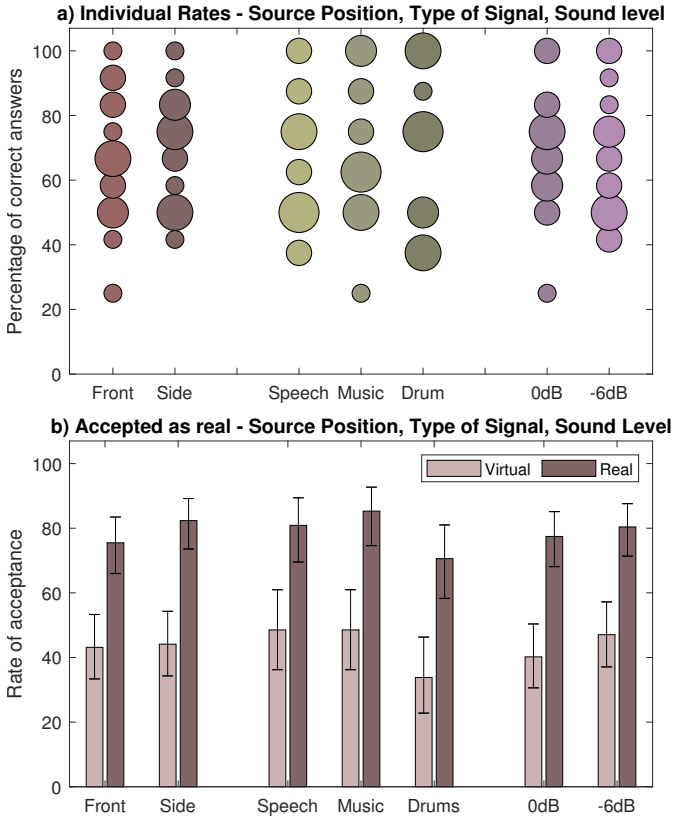


Figure 4.20.: a) The individual rates of correct assignments were not significantly influence by the position of the sound source, the type of source signal or the sound level - b) No significant influence of source position, type of signal and sound level on the acceptance of the real and the virtual sound source in Part II could be observed in this experiment.

For the central question in this experiment, the percentage of correct answers gives only limited insight. So instead, it is of interest to analyze the acceptance as *real*. A separate analysis of the individual amount of correct answers for virtual and real scenes for each condition was not feasible. This is because the sample size per person is already relatively small for all of them together.

However, a pooled inspection is possible. Fig. 4.20 visualizes the pooled rate of scenes accepted as real per condition separated by virtual and real reproduction for the whole pool of participants. Again, only the results of part II of the experiment were taken into account. In addition to the bars indicating the *percent correct* for each condition, the Clopper-Pearson Confidence Intervals (CIs) [267] are shown. The virtual sources were accepted as *real* significantly less often than the real source for each condition. There is an overlap of the CIs for the correct identification of the real scenes (SDT: Correct Rejections), and also, the percentage of virtual scenes accepted as real (SDT: Misses) does not vary significantly with source position, type of signal, or sound level.

In summary, neither the source position nor the level or type of signal significantly impacted the plausibility. This is especially interesting regarding the source position, considering that with the source position, the listener's motion relative to the loudspeaker was different. For the frontal sound source, the subjects could walk towards and away from it. For the position right of the translation line, the participants could walk past the front of the loudspeaker. The directivity of the sound source has a substantial impact on the progress of the direct sound. These difference to the frontal condition did not yield different plausibility (as the agreement with the tuned internal reference).

Source position, type of signal, and sound level

Fig. 4.20 provides an overview of the rates of correct answers for the source position, the type of signal, and the sound level. Only part II of the experiment is considered for this analysis.
The first graph visualizes the subjects' individual rates of correct answers within the test. For each source position and each sound level condition, the total amount of test cases per person was twelve, six real and six virtual. According to the paired t-test, the number of correct answers for the source in the front ($M = 7.94$, $SD = 2.38$) does not differ significantly, $t(16) = 1.0$, $p > .3$, from the percentage of correct answers for the lateral source position ($M = 8.29$, $SD = 2.02$). The percentage of correct answers for the 0 dB sound level ($M = 8.24$, $SD = 2.22$) was not significantly different, $t(16) = 0.57$, $p > .3$, from that for the -6 dB ($M = 8.0$, $SD = 2.29$).
For each type of signal, each participant rated eight scenes in part II, four virtual and four real. The individual rates of correct answers for speech ($M = 5.29$, $SD = 1.61$), music ($M = 5.47$, $SD = 1.70$) or snare ($M = 5.47$, $SD = 1.94$) were not significantly different from each other, $t(16) < 0.5$, $p > .6$ for all three combinations. In summary, the position of the sound source, the type of signal, and the sound level did not significantly influence the percentage of correct answers.

4.4.3. Discussion and Conclusions

In this experiment, the plausibility of an auditory AR illusion created over headphones for a position-dynamic exploration by the listener was evaluated with regard to the pure internal reference on the one hand and with regard to an internal reference that was tuned by including the real counterpart of the test scenes on the other hand.

Plausibility of position-dynamic AAR realization for interactive walking and revealing cues

When the real test scenes were included as hidden references, experienced listeners could identify binaural auralization quite confidently, and inexperienced listeners did not predominantly accept the virtual reproduction as real anymore as in part I.

One of the primary cues to identify the auralization was the audible difference in case the listener turned his back toward the source. Distance perception, externalization, and timbre were affected. In none of the previous studies was such an effect documented. Brinkmann et al. [29] tested the authenticity for source directions of 0° and 90°, allowing a head rotation of ±34°. The study was conducted with extra-aural headphones. Lindau and Weinzierl worked with STAX headphones and allowed a head rotation of ±80°. Pike et al. [262] also used STAX and provided a system capable

of a full 360° reproduction, but instructed their participants to move only their heads but keep their torso still. The case of the source in the back has not received any attention so far. This means that our study is also the first we know to investigate plausibility with regard to the tuned internal reference for dynamic binaural synthesis with "true 360°". It is hard to tell whether the observed effect in the back is unique in the system used for this study or whether it is a general phenomenon. In none of the previous studies AKG K1000 headphones were used. Satongar et al. [261] showed that the passive influence of headphones could cause spectral distortions, affect the effective interaural time difference and reduce localization accuracy. However, their study did not consider the AKG K1000. Measurements of the physical effect of AKG K1000 headphones by Pörschmann et al. [268] and Schneiderwind et al. [278] indicate that these might contribute to such audible effects.

Another cue was the slight instability of the source position during quick head rotation. Similar observations were reported in an earlier study by Lindau et al. [214] testing an early-stage system and by Pike et al. [262]. This audible effect could be due to non-individualized ITDs or a non-optimal delay of about 100 ms in the motion-related updating of the BRIR filters. These aspects must be improved to achieve an authentic or plausible (with regard to tuned internal reference) reproduction.

Five subjects mentioned that they localized the sound source in the head before moving. They assigned this experience to the binaural simulation. However, in-the-head-localization can occur in real sound fields as well [269]. It is questionable whether this is a reliable cue for identifying virtual sound sources. Still, it may occur more often or be more pronounced in binaural reproduction.

Four participants stated that the change of level during walking was a helpful cue for them. They reported that the level would change not enough or too much over certain sections of the translation line. These effects were also reported in previous experiments on the plausibility of an approaching motion [81, 324] (Sec. 4.1 and 4.2). Several untrained listeners were surprised about the progress of the sound level in the measured scenario. They rated the manipulated version of the scene as more plausible because the level change was closer to what they expected. This may also be a case of an inaccurate or wrong internal reference. In fact, also in the present experiment, this cue was only reported by untrained listeners.

Three participants reported a confusing localization which includes increased elevation (higher than the visual source) and reduced sharpness in the image of the sound sources. An increased elevation in the localization is a common artifact in the binaural simulation with non-individual BRIRs. This is likely a reliable cue revealing the simulation for some people. An increased blurriness might result from reproduction with generic BRIRs as well.

Furthermore, two participants perceived differences in the timbre and stated that the simulation has less strength in the low frequencies. The stimuli were limited to a frequency range between 150 Hz and 16 kHz for both reproduction methods. Deviating timbre might be caused by the non-individual BRIRs and a non-individual headphone compensation.

Two people reported an increased apparent source width. This usually occurs with an increase in reverberant energy. However, these reports may be connected to the reduced sharpness of the sound image when listening to a real sound source while wearing headphones.

This experiment was the first to consider position-dynamic binaural synthesis and their corresponding real version of the sound field in a test scenario with interactive self-translation of the listener. Furthermore, this study was the first to consider a true 360° experience when studying the discriminability of the auditory illusion from its real version.

Most of the cues reported as helpful for identifying the virtual version were unrelated to translation. For example, four untrained subjects mentioned that the sound level would exhibit unexpected progress during walking. Similar statements were given in the previous experiment by [324] (Sec. 4.2) for the measured scene by participants who rated another artificial scene with a more distinct change of the level as plausible. This judgment may result from an inaccurate or wrong internal reference. 13 of the 17 subjects in the present experiment did not mention any translation-related cues. Thus, the present realization of the translation did not cause substantial effects revealing the binaural auralization. However, without the additional freedom of motion in this test scenario, observing the

unnatural impression of the sources in the back may not have been possible. In addition, it is interesting noticing that no significant differences between the cases of walking past and towards/away from the loudspeaker were observed.

Influence of the availability of the real version - Pure versus tuned internal reference

Creating a test design investigating the influence of the availability of real versions of the sound source on the estimated plausibility is not straightforward. It has to be taken into account that the test without real reproduction always had to be conducted first and without any training. Especially for inexperienced listeners, it is likely that it takes a while to identify helpful cues and establish strategies for efficient exploration. Such effects could not be eliminated with the given test design. Then again, identifying helpful cues revealing the virtual scene may be easier when a real scene is presented in between. For the progress of the share of correct answers over the trials in the tested order, a regression analysis was conducted. This analysis is independent of the actual test condition. Both parts of the experiment were analyzed separately. The hypothesis that the regression coefficient is zero could not be rejected ($p >$0.6 in both cases). This indicates a flat "learning curve" with no trend or evident increase in the number of correct answers during the experiment. Consequently, it is reasonable to neglect the effects of training or getting used to the task for conclusions based on the submitted answers.

Another influence might be an expectation of the participants that real and virtual test scenes may be equally distributed in the test sample or a least a certain minimum amount of both options is included. This might have an effect if, in part I, subjects are not sure of the answer and become irritated by having the impression of repeatedly listening to the virtual version. In these cases, subjects might answer *real*, although they tend to answer *virtual*. However, this is only an issue if a subject cannot confidently identify the virtual reproduction. In contrast, at least several inexperienced listeners often answered with *real*. They did not mind giving the same reply repeatedly.
To minimize this issue in part I, after 12 virtual scenes, 12 real scenes should be tested in addition. Then part II with the same scenes in randomized order could follow. In that setup, the number of correct answers for the 12 real scenes in a row would be affected by the same psychological bias. As a result, the percentage of correct answers and, thus, the acceptance rate would be reduced. Comparing the results of this part to the purely virtual part in terms of the paired t-test or calculating the sensitivity index would be less critical than comparing it to the results of the real scenes in part with the randomized order. We decided not to include such a part in the experiment because the test was quite long already. Instead, we chose to use a more critical evaluation by comparing the results of part I to the real scenes in part II. We assume that this decision does not affect the main findings of this experiment.

The results of this experiment suggest that including the real version of the scenes affects the listener's capability to identify the simulation. The test design with randomized order of different signals, source positions, and sound levels minimized the options for a direct comparison between a virtual scene and its real counterpart. Thus, we can conclude that the test design influences internal reference, which is fundamental for evaluating plausibility.
The fact that including the real version affects the estimated plausibility and reduces the acceptance of virtual imitation is not surprising. It is known from other test methods that the choice of test items influences the test results for the single items and that including a (hidden) reference representing the best possible quality facilitates critical testing as discussed, for example, by Zielinski et al. [270]. The observations indicate that discriminating between different kinds of plausibility may be of interest in the future. On the one hand, the plausibility that measures the agreement with the listener's pure internal reference will be of interest, e.g., in the case of fictive scenes. On the other hand, the plausibility that measures the agreement with an internal reference tuned by listening to a real version of the scene will allow for a more critical evaluation.

In augmented acoustic reality, the real environment is always present and will provide a reference for a virtual acoustic element. For evaluating its quality, it is crucial to consider the influence of the elements and properties of the real acoustic environment. Authenticity is evaluated in a direct comparison of a virtual and a real scene and is even more sensitive.

How should the plausibility of AAR be evaluated in the future?

This study considers an AAR scene, which contains one primary sound source besides the typical quiet background sound in everyday environments like the chosen seminar room. The participants experienced the room with its acoustic behavior when they entered the room, walked to the test setup, talked to the test conductor, and got the introduction. This is likely to cause certain expectations regarding how the reproduction of the loudspeaker standing in the room should sound. However, more complex scenes which contain a variety of real and virtual sound sources are more interesting and more common for application scenarios of AAR. In such scenarios, there is usually no option to listen to precisely the real version of the virtual sound element at exactly the same position. Instead, the real sound sources of the actual acoustic environment are available among the virtual contents and serve as an external reference to some extent. Wirler et al. [217] already showed that the scene complexity affects the plausibility evaluations. The results of our study suggest that an available real equivalent to the virtual sound object will have a tuning effect on the internal reference. Further studies are necessary to improve the understanding of a listener's internal reference and its interrelation with different types of external reference. This is especially interesting in the case of fictional content, how their perception and acceptance are influenced by the other real and virtual elements of the given scenario.

Evaluating plausibility with regard to the pure internal reference has the advantage that considering the headphones in the BRIR measurement is not required. In this experiment, headphones had to be taken into account to focus the investigation on the test method and avoid changing more than the primary variable among the test conditions. However, apart from the significant differences between both test methods, we observed that the primary cue for identifying the virtual reproduction among the real scenes was probably caused by the shadowing effect of the headphones. This raises the question of whether the significant differences in plausibility hold if the evaluation with respect to the pure internal reference was conducted with BRIRs neglecting the occlusion effect. With regard to the desired ecological validity of test methods in general, both methods are of equal interest. For AR, the listener will always have to wear a listening device. Despite all attempts to create a transparent headphone experience, perfect transparency has not been achieved yet. Then again, the overall goal is to create auditory illusions that appear in the real world without the slight influences of any headphones.

Summary

The experiment presented in this article was conducted to evaluate the plausibility of walk-through scenarios with position-dynamic binaural synthesis using a state-of-the-art system. The realization is based on BRIR filters measured with a Kemar head-and-torso-simulator wearing AKG K1000 headphones in the room and at the positions where the psychoacoustic experiment took place. The subjects could see two loudspeakers in the room, and in each scene, one reproduced sound virtually or in reality. The subjects could either walk past the sound source or towards and away from it in different test cases. Head rotation and self-rotation were possible at all times. The subjects had to determine whether they heard the real reproduction or its binaural simulation in each trial. Dry male speech, a snare drum sample, and music in terms of a pop song were investigated. The experiment was divided into two parts. In part I, the plausibility was evaluated with regard to the subject's pure internal reference without the option to listen to a corresponding real version of the simulated sound

fields. In part II, the approach of discriminating the binaural auralization from the corresponding real sound fields as proposed by Lindau and Weinzierl [211] was applied to binaural walk-through scenarios with a true 360°-experience for the first time. Including real sound scenes as test items come with some challenges and limitations. On the one hand, the method can only consider the real scene as it is perceived through headphones or hearables. On the other hand, these effects have to be considered in creating auditory illusions, for example, by measuring an extra set of BRIR measurements, including the hearing device of interest. Moreover, the method can only consider contents where a corresponding real version is available. In three earlier studies, the given system has repeatedly been rated as plausible in an evaluation without any real scene. If no real scene is included, it is not necessary to take the occlusion or shadowing effects of the headphones into account in creating the virtual content. Thus, there is no optimal evaluation method. In addition to the previous experiments, the present study evaluates the plausibility in a Yes/No paradigm with and without including the real versions of the simulated scenes as hidden references.
With the given AAR system, the inexperienced listeners accepted the virtual version as real in most cases, in part I, when the real scenes were unavailable. In this case, even the experienced listeners could not confidently identify the presentation as a simulation. In contrast, in part II, when the real versions were available in the test, experienced listeners could detect the simulation quite confidently, while inexperienced listeners at least increasingly doubted the realness in the case of the virtual version. Source position, type of walking motion relative to the source, type of source signal, and sound level did not significantly influence the observations. There were two primary cues revealing the virtual reproduction. In the listener's back, the sound source exhibited an unnatural appearance caused by the headphones' presence. In addition, the participants reported slight instabilities of the sound source during head rotation which was probably caused by the lack of individualization and maybe the non-optimal system latency of nearly 100 ms.

Conclusion

The results of the presented study indicate that the system under test can induce a plausible illusion for inexperienced listeners. However, the system fails to deliver a plausible illusion for experienced listeners in general and all listeners if they had the chance to listen to the real counterpart of the sound field. The primary cues affecting plausibility are not caused by the increased freedom of motion of this AAR setup but rather introduced by the presence of the headphones and the lack of individualization. As expected, the results show that the availability of a real counterpart tunes the internal reference and leads to a more critical evaluation of plausibility. On the one hand, this suggests that the presence of similar real sound objects in an AR scenario may also affect the plausibility of virtual content. On the other hand, this evaluation method requests to consider the occlusion effect of the headphones also in the synthesis of the virtual content. This reduces the overall quality of the AR reproduction and limits the ecological validity of this test approach. However, the fact that perfectly transparent headphones are not available remains a challenge for realizing AR systems. Especially for motion in 6DOF, the knowledge about this influence on the perception of real sound sources is still surprisingly low. Under these test conditions and compared to these effects, potential imperfections of the position-dynamic binaural synthesis used in the system under test did not appear critical for the plausibility of the AAR realization.

4.5. Experiment V: Effect of manipulations in the early reflections on plausibility in an augmented reality scenario

This section is based on a journal article (preprint) "Effect of impaired early reflection patterns on plausibility and similarity of position-dynamic binaural AR audio" [314] that I submitted for publication.

The two experiments described in Sections 4.1 and 4.2 showed that creating a plausible interactive auralization for walking towards and away from the loudspeaker is possible without a detailed adjustment of the reverberation to the position of the listener. Even keeping the reverberation constant right after the direct sound was perceived as plausible by most participants for the tested translation path in both rooms. However, only a small range of listening positions in front of the loudspeaker was taken into account, and moving sideways, past or behind the sound source was not possible. Moreover, the participants rated the plausibility of the various test scenes without seeing the room or an object representing the sound source, but only a neutral spatial grid environment. In addition, the plausibility was evaluated in a single-stimulus test design similar to the Absolute Category Rating specified in Rec. ITU-T P.800.2 (07/2016) [271]. In such a test design, the listener relies on his internal reference because a direct comparison between the scenes or the real counterpart is impossible. The internal reference is individual and can also be vague or even wrong. Furthermore, the chosen types of manipulation are also limited in their ecological validity with regard to the various approaches of simulation, auralization, interpolation, and extrapolation that have been documented in the literature already. Sec. 2.6 provides a short overview.

To overcome these limitations, the final experiment of this thesis was designed. On the one hand, the test approach is changed from (Multiple Attribute) Absolute Category Rating to a (slightly modified) Paired Comparison. Furthermore, this time, the experiment is based on an AR scenario, where the participants could see the actual room and the loudspeakers whose sound emission should be imitated. This demands audiovisual plausibility.

Moreover, in addition to the loudspeaker in front, the second loudspeaker position on the right side of the translation line was taken into account. Furthermore, to consider the effects of the source directivity, both loudspeakers were turned by 180°, and the whole measurement was repeated. Thus, two extreme cases of source directivity were included in the experiment. Now, the listener can move towards, past, behind, and away from the virtual sound source, and scenarios with low direct sound energy are included. This study aims to explore the perceptual effects of implausible early reflection patterns and determine their impact on the plausibility of spatial auditory illusion.

4.5.1. Measurement: Indirect irradiation scenario in position-dynamic binaural audio

The measurement was again conducted in the same seminar room with a size of 9.9 m×4.7 m×3.1 m, a volume V = 144.2 m^3, a reverberation time T_{30} = 0.98 s and a Schroeder frequency of about 165 Hz. Fig. 4.21 shows the room. The two loudspeakers Genelec 1030A remained in the same positions but were rotated by 180°, facing away from the translation line to create an *indirect irradiation scenario*. The line for listener translation remains in the same position. A Kemar 45BA head-and-torso-simulator (HATS) was subsequently placed at nine positions in equivalent intervals of 25 cm along the translation line. Because the furniture had been slightly rearranged, the *direct irradiation scenario*, equivalent to the measurement described in Sec. 4.2.1, was also measured again. The arrangement is illustrated in Fig. 4.22.

In addition, the microphone array shown in Fig. 4.23 was placed at each of the nine positions to capture directional room impulse responses. These are of interest to analyze the spatio-temporal pattern of early reflections.

Figure 4.21.: (A) Setup in the seminar room, forming the *direct reproduction scenario*, where the loudspeakers are turned towards the translation line - (B) Loudspeakers turned by 180° to create an *indirect reproduction scenario*.

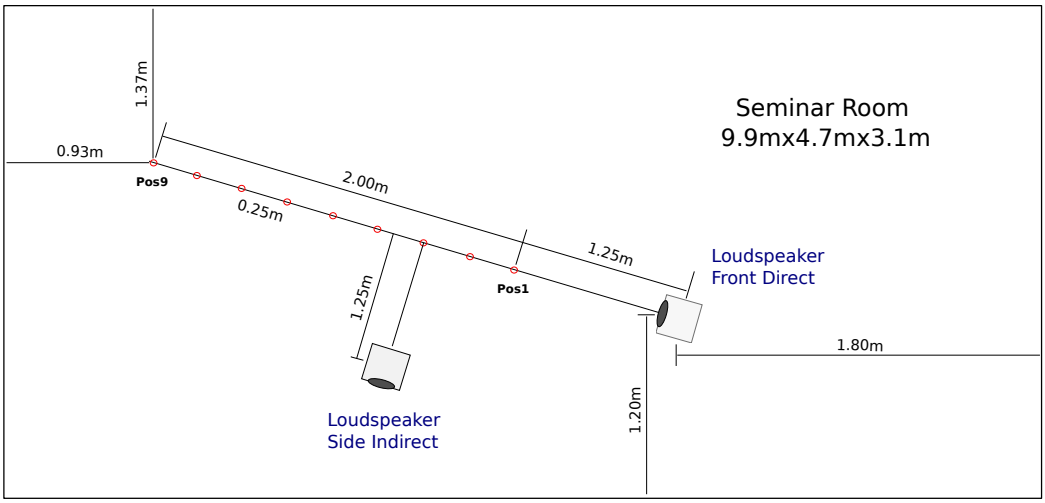


Figure 4.22.: Illustration of the setup in the seminar room with the two loudspeaker positions and translation line, consisting of nine measurement positions, with *Pos1* being the closest to the frontal loudspeaker and *Pos9* the farthest.

Figure 4.23.: Microphone array to measure directional room impulse responses for analyzing the spatio-temporal pattern of early reflections at the nine listening positions along the translation line.

Mixing Time Predictions for the given room

Lindau et al. [150] proposed models to predict the average perceptual mixing time t_{mp50} to achieve a transparent simulation for 50% of the listeners and t_{mp95} for the majority. For the chosen seminar room with V = 144.2 m^3 the estimated perceptual mixing time would be

$$t_{mp95} = 0.0117\,V + 50.1\,ms = 51.79\,ms$$
$$t_{mp50} = 0.58\,\sqrt{V} + 21.2\,ms = 28.16\,ms.$$

The room contains some furniture, as shown in Fig. 4.21. Thus, the volume of interest may be smaller, but the effect on the estimated perceptual mixing time is minimal in small rooms.
In addition, Lindau et al. [150] proposed the prediction with signal-based models considering t_{mix_Abel} according to Abel and Huang [272]. AKtools [273] provide a script to estimate t_{mix_Abel} from a measured RIR. At each of the nine positions along the line, RIRs were measured for the four different arrangements of the sound source. The script was applied to all of these RIRs and the resulting values ranged from 30 ms to 85 ms with a mean of t_{mix_Abel} = 49.4 ms. Consequently, t_{mp} can be predicted with

$$t_{mp95} = 1.8 \cdot t_{mix\_Abel} - 38\,ms = 50.9\,ms$$
$$t_{mp50} = 0.8 \cdot t_{mix\_Abel} - 8\,ms = 31.5\,ms.$$

Physical analysis of early reflections

In the Front-Direct scenario, the six strongest reflections arrived at the closest listening position (Pos1) about 6 to 28 ms after the direct sound with levels from -20.8 to -15.3 dB relative to the direct sound and at the farthest position (Pos9) 4 to 20 ms after the direct sound with levels of -10 to -5 dB. For the Front-Indirect scenario, the reflections arrived between 23 to 40 ms with levels of 0.65 to 7.4 dB. For the Side-direct case, the six strongest reflections arrive at Pos1 again about 6 to 28 ms after the direct sound with a level of -19.6 to -12.8 dB, and at Pos9 6 to 36 ms after the direct sound with levels of -11.4 to -5.7 dB. For the Side-Indirect case, the six strongest reflections arrived within a similar time-range like for the direct case, but with 0.7 to 9.2 dB at Pos1 and 3.8 to 13.6 dB were again stronger than the direct sound. Tab. A.3 provides a detailed overview.

Fig. A.5 visualizes the predictions of perceptually salient reflections according to Brinkmann et al. [80]. Close to the front of the loudspeaker (graph (A)), only very few early reflections seem to be relevant. Two meters further away (graph (B)), already many more reflections have to taken into account according to this model. Turning the loudspeaker by 180° (graph (C)) leads to a further substantial increase of the number of salient reflections. For the loudspeaker on the side of the translation line, similar relations were observed.

4.5.2. Methodology

For the direct reproduction scenario, it is of interest to confirm the findings from the previous experiment regarding the constant reverberation with the new test method and to challenge this psychoacoustic effect by switch the left and right channel after the direct sound to create another obviously implausible spatiotemporal pattern of early reflections.

The indirect reproduction is used to generally challenge approaches for position dynamic auralizations, however, in particular the constant reverberation.

Manipulations of the measured BRIR set and pre-test

The manipulation started with automatic detection of the direct sound peak and applied it to the left and right BRIR channels for all 90 directions captured at one position. The temporal average was calculated and considered as the time stamp of the direct sound of each of the 90 BRIRs. Visual inspection revealed that 55 samples ($f_s = 48$ kHz) after this time stamp, the direct sound was over for all directions in both channels, and the early reflections started just after that. If required for a test case, the reverberation part is split from the direct sound at these 55 samples after the mean peak time of direct sound with a cross-fade based on a complete overlap of the two half-Hann-windows with an overall duration of 32 samples.

LR - Left-Right shift in reverberation - The first, very basic manipulation was a shift of the left and right channel, starting after the direct sound. For each of 9×90 BRIRs, direct sound and reverberation were split as described above. This test case is of interest to investigate the sensitivity to an implausible spatiotemporal pattern of early reflections with regard to the visible environment.

LR50 - Left-Right shift after 50 ms - The left and right channels were exchanged starting from 50 ms after the mean peak of the direct sound using the described cross-fade approach.

CRev - Constant Reverb - The reverberation part of all BRIRs measured along the line was replaced with the reverberation measured at Pos 1 for a head orientation of 0° (facing the frontal loudspeaker). For the indirect reproduction and the side loudspeaker scenarios, the reverberation was also taken from Pos 1, 0°azimuth, but from the BRIR measured for the respective source constellation.

CRev50 - Constant Reverb after 50 ms - The same procedure as for CRev was realized, starting 50 ms after the mean peak of the direct sound.

GMT11 - Global Mixing Time - In earlier studies, the mixing time was only considered for one distance at a time. The time of transition between the early and late reverberation part was usually determined with respect to the direct sound (peak). However, in a sound field, it may be more interesting to have one global mixing time for the whole sound field that only has been sampled at the nine measurement positions. Such an approach demands a global time reference which is the same for all positions. In this specific case, we picked the point of time when the sound source starts to play. After a short traveling time, the direct sound will successively arrive at each of the measurement positions and spread to the remainder of the room. A global mixing time of 11 ms (GMT11) was the smallest value possible, considering that sound takes about 10 ms from the frontal loudspeaker to the farthest measurement position. Consequently, at the first position, the transition will take place 8 ms after the direct sound for the direct reproduction, 7 ms for the indirect case, at the farthest direct 2 ms, indirect 1 ms, if the frontal loudspeaker position is chosen. The duration from the direct sound to the transition will be between these values for the lateral loudspeaker. The transition between the early and late part of the BRIRs was again realized with the short Hann window. Fig. 4.24 visualizes the differences between the CR and GMT approaches to use the perceptual mixing time.

GMT60 - Global Mixing Time - A global mixing time of 60 ms was included as an equivalent to CRev50, following the estimated tmp95 according to [150].

GMT40 - Global Mixing Time - A global mixing time of 40 ms was included as an equivalent to tmp50 - as suggested by Lindau et al. [150].

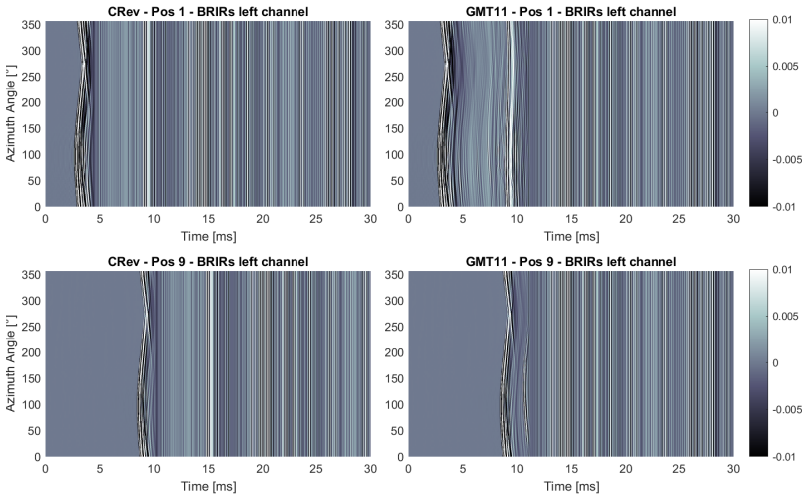


Figure 4.24.: BRIRs (left channel) for the different angles at Pos1 and Pos9 for manipulations CRev and GMT11. For CRev the transition time is related to the position-dependent time of arrival of the direct sound, GMT considers a global time reference - the start of the sound emission.

Anchor - DRR -4.5dB - For the direct reproduction scenario, the pre-test indicated that all manipulated versions are very similar to the original. It might be confusing for a participant if a block of test scenes that all sound very similar have to be rated. Therefore, an anchor condition was introduced that is known to be not as plausible as the original. Reducing the reverberant sound energy by 4.5 dB to increase the DRR by 4.5 dB received significantly lower ratings for audiovisual plausibility in a previous study [327]. Listeners reported perceiving the virtual sound source in front of the loudspeaker, which indicates a lack of audiovisual coherence. Therefore, this seemed a reasonable choice as an anchor condition. The reverberant part was separated from the direct sound described above and then decreased by -4.5 dB.

Test setup

Again, the experiment was conducted in the same positions in the same room where the BRIR measurements were conducted. The 2 m-translation line was marked in red on the floor. The listener was invited to walk along this line and arbitrarily turn in all directions to explore the scene. Both loudspeakers used in the measurement were kept in their positions to serve as visual references in the experiment.

In this experiment, the goal was to evaluate the audiovisual plausibility in the chosen seminar room. Consequently, this time the setup was designed to minimize the limitations and influences on the visual and audible impression of the real environment. Extra-aural headphones AKG K1000 were chosen to reduce the influence of their presence on the perception of the real acoustic environment. The choice is explained and argued in Sec. 3.4.2.

An HTC Vive-tracker was attached to the headphones to track the motion of the listener's head. PyBinSim [317] was used to handle the BRIR filter selection according to the current position and orientation of the listener. The partitioned convolution was realized with a blocksize of 256 samples at a sampling rate of 48 kHz. In addition, a headphone compensation filter was applied according to the least-squares-approached described by Schärer and Lindau [53].

The participant was provided with a tablet with a graphical user interface containing four buttons. With "A" and "B," the subject could switch between the two scenes of the test pair. "Pause" paused the audio reproduction, and with "Next," the next test pair was loaded. The GUI was a web interface, and the communication with pybinsim was realized with the open-source AVRate Voyager framework [274].

Figure 4.25.: Setup for the listening experiment: Participant wears AKG K1000 headphones with an HTC Vive Tracker attached to it and walks along the translation line marked in red on the floor, switching between test scenes A and B via a graphical user interface provided on a tablet PC.

Test method

Lindau and Weinzierl [211] proposed to evaluate plausibility in a yes/no paradigm to determine whether simulations can be identified among real versions of the sound field. Considering real references requires taking the shadowing effects of the headphones into account in creating the virtual content. In the previous studies [81, 324], plausibility was evaluated in a single-stimulus or better single-scene paradigm without considering a real reference. However, in typical AR/MR scenarios, additional real sound sources are likely to be present. The findings of Neidhardt and Zerlik [275] as well as Wirler et al. [217] suggest that the presence of real sound sources may have a tuning effect on the internal reference even if it is not exactly the real counterpart of the virtual sound object.

To find a good compromise, in this study, plausibility is evaluated in a paired comparison of only virtual versions with the scene created from the fully measured BRIR set serving as the reference. With the given AR system, the user can see the real loudspeakers. The plausibility is likely to be affected by noticeable constraints of the audiovisual coherence. Therefore, we also speak of *audiovisual plausibility*.

Additionally, it is interesting whether the manipulations cause audible differences with respect to the reference. Thus, *similarity* was added to the evaluation.

Assuming that the original measured data set would create a perfectly plausible illusion for each participant is not feasible. On the one hand, the setup contains only non-individualized BRIRs. On the other hand, variations in elevation are not considered in this implementation because it was desired to include a fully measured data set, which is very cumbersome if different elevation angles should be included as well. However, such a setup does not correctly account for the individual body height of the different participants. Therefore, the subjects were asked to evaluate the plausibility of both items in the evaluation pair separately on the scale shown in Fig. 4.26. The 4-point scale used in the previous studies was extended to a 6-point scale to allow for better differentiation, which appeared desirable in past experiments.

In addition, the similarity of the two scenes A and B in the test pair had to be rated on a scale from 1 to 7 shown in Fig. 4.27. A rating of 1 indicated that both scenes sound the same and 7 indicating that both are very different from each other.

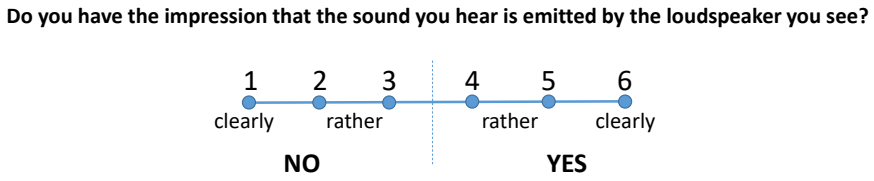


Figure 4.26.: Scale for evaluating plausibility.

Figure 4.27.: Scale to rate similarity between scenes A and B.

Test scene pairs

With the goal to compare each condition of interest to the measured reference and considering all four loudspeaker scenarios, it was challenging to keep the test duration reasonably short. Furthermore, all conditions should be tested twice to analyze individual consistency. As a consequence, only a few conditions could be selected. Finally, each condition was paired up with the measured reference scene.

In addition, for each loudspeaker scenario an anchor pair was included which did not contain the original measurement. These pairs were selected so that both items differed from the measurement. The goal was to remind the subjects that the measured reference is not always part of the test pair and that they should remain critical with each test scene and start the exploration and evaluation of each scene from scratch. In the direct scenario, the anchor scene with the reduced reverberant energy was used twice to build a pair. For the indirect scenario, LR-CRev was chosen as the anchor pair because the pre-test indicated audible deviations from the original, likely to be perceived as less plausible. Table 4.6 provides an overview of the final selection of pairs.

Table 4.6.: Test pairs for the experiment

Front Direct (FD)	Side Direct (SD)	Front Indirect (FI)	Side Indirect (SI)
Orig - Orig	Orig - Orig	Orig - Orig	Orig - Orig
Orig - LR	Orig - LR	Orig - LR	Orig - LR
		Orig - LR50	Orig - LR50
Orig - CRev	Orig - CRev	Orig - CRev	Orig - CRev
		Orig - CRev50	Orig - CRev50
Orig - GMT11	Orig - GMT11	Orig - GMT40	Orig - GMT40
		Orig - GMT60	Orig - GMT60
Orig - Anchor	Orig - Anchor		
Anchor pair	Anchor pair	LR - CRev	LR - CRev

Test stimulus

In this experiment, only male speech was used as a test stimulus. It is known from earlier studies that, for example, noise is a more critical stimulus regarding the continuity of the reproduction with the limited angular and positional resolution [256], [81, 326]. A different reproduction setup or

rendering approach is required to investigate the same research questions for noise-like signals or we also wanted to avoid noise burst to ensure apparent continuity of the sound field. Broadband noise is likely to be more critical with regard to similarity. However, for plausibility it is hard to say whether noise would lead to changes. Most people may have a sharper internal reference for speech than for noise being reproduced by a loudspeaker which might make noise less critical for this test case.

Test procedure

The four source position and orientation scenarios (FD, SD, FI, SI) were tested in four individual test blocks. Every participant had to complete each of the test blocks once. The experiment was conducted in two sessions on two different days. Each session consisted of two blocks, including front and side, as well as direct and indirect cases (e.g., FI-SD or SI-FD). The order and combination of the test blocks were pseudo-randomized to achieve a nearly equal distribution over the eight possibilities.

To explore a pair of test scenes, the subject could switch between scenes A and B while moving arbitrarily on the given translation line. The subjects were instructed to avoid sudden movements because the headphones did not remain stable on the head. Similarity and plausibility had to be evaluated within the same comparison procedure in no specific order. After announcing the three ratings (two plausibility ratings and one for similarity), the next pair of test scenes could be loaded. The order of the test pairs was randomized individually within each test block.

Participants

18 volunteers, 7 females, 11 males, participated in the experiment. They had an average age of 31.8 years with a standard deviation of $\pm$ 5.2 years. Two people reported experiencing minor tinnitus sounds, and one reported a lack of sleep. Otherwise, all participants reported having normal hearing abilities without any known impairments. Eight of the 18 participants are very experienced listeners of binaural audio, as it is part of their professional lives. Five had also taken several experiments in this field and partly conducted one or two themselves in terms of student projects. Four were reported to have only rarely (1-3 times) experienced dynamic binaural audio reproduction before. For none of the subjects, it was the first time to participate in a psychoacoustic experiment focused on spatial auditory illusions created over headphones.

4.5.3. Results

Each participant rated each pair of test scenes twice. Overall, each item was rated 36 times for similarity to the measured original and 36 times for plausibility in a paired comparison to the measured original. The results are visualized in Fig. 4.28, 4.29 and 4.30.

A Shapiro-Wilk test indicated that the plausibility and similarity ratings for the individual test case are not normally distributed.

Plausibility - A Quade test was conducted as an omnibus test for the group of test conditions (without the anchors) separately for each of the four source constellations. It indicated significant differences among each of the test groups. Therefore, a paired, two-sided Wilcoxon signed-rank test was used to test the hypothesis that the differences between the matched plausibility ratings of the compared test conditions comes from a distribution whose median is zero. The significance level was corrected for multiple testing using the modified false discovery rate (mFDR) method (also called B-Y method), because it is less conservative than the Bonferroni correction [243].

For the direct reproduction scenario, only the comparison between the original and the anchor condition yields significant differences in plausibility with $p < 0.001$ for both loudspeaker positions. For the indirect scenario, significant plausibility differences were found for the comparisons *Orig-LR* ($p < 0.001$), *Orig-CRev* ($p < 0.001$) and *LR-CRev* ($p < 0.001$) for both loudspeaker conditions. *LR-CRev* was added to serve as an anchor pair. Thus, only the indirect reproduction scenario produced cases where the plausibility was significantly affected by the manipulation of the ER. This is very interesting because *LR*, *CRev*, and *GMT11* are quite severe modifications, in particular with regard to the geometry of the room and the reproduction setup. A modification of the late reverberation (50 ms after the direct sound or 60 ms after the start of the sound emission) did not affect the plausibility in any of the test cases. Moreover, also *GMT40* with the slightly earlier start of the modification in correspondence to the predicted t_{mp50} in the indirect scenario did not exhibit a degradation in plausibility.

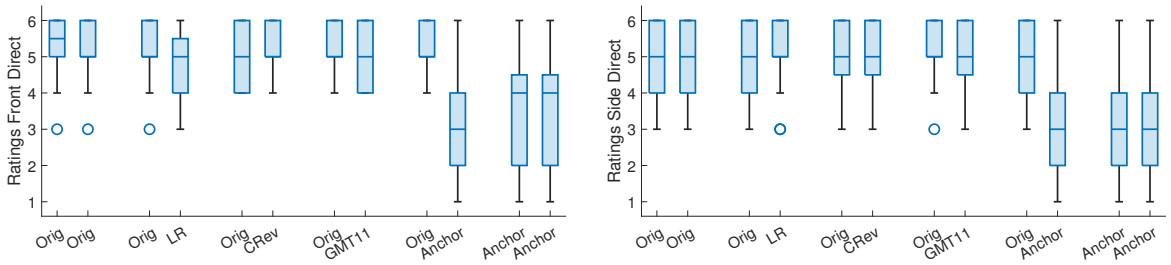


Figure 4.28.: Results for the audiovisual plausibility in the direct reproduction scenario

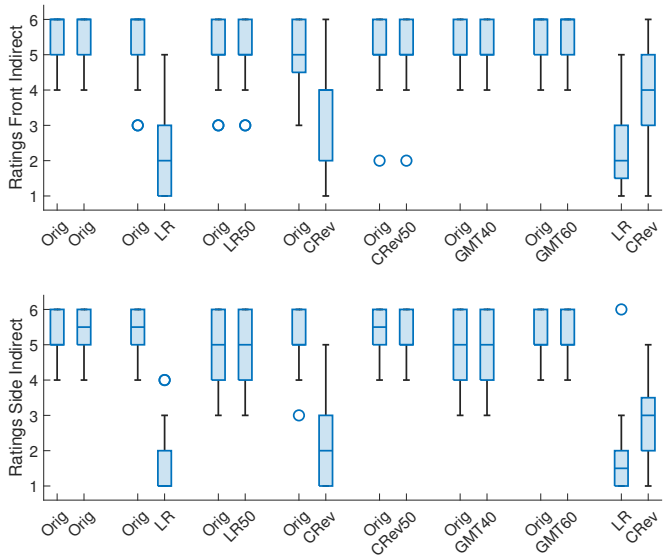


Figure 4.29.: Results for the audiovisual plausibility in the indirect reproduction scenario

For the originally measured data set, perfect plausibility could not be assumed. The lack of individualization and of considering elevation changes caused by different body heights are examples of limiting factors. Some participants reported an increased elevation of the source image, affecting the audiovisual coherence. In the earlier studies, a few participants preferred a considerably modified scene in terms of plausibility. Such effects may be caused by an inaccurate internal reference or phenomena like hyperrealism. However, in the present study, none of the modified BRIR sets was rated significantly more plausible than the reference. Therefore, choosing an original measurement as a reference seems reasonable.

The observations by Neidhardt and Zerlik [275] and Wirler et al. [217] indicate that a real reference

can affect the plausibility of the scene under test. Could the scene under test also affect the plausibility of the reference? Additional paired, two-way Wilcoxon signed-rank tests investigated whether the ratings for *Orig*-scene depend on the condition it was directly compared to. For both direct reproduction cases, no significant impact of the paired scene was found (all $p > 0.18$). However, for the Front-Indirect-case, the plausibility ratings of the original scene were significantly affected when compared to CRev ($p = 0.014$) and CRev50 ($p = 0.013$) instead of comparing it to itself. The mFDR-corrected significance level is 0.0193 for multiple testing of seven pairs.

For conditions LR50 ($p = 0.05$) and GMT40 ($p = 0.039$), the p-value is only slightly larger than the corrected significance level, but when compared to LR ($p = 0.40$) and GMT60 ($p = 1.0$), the reference was rated just as good as when it was compared to itself. For LR, we know from the results that it sounds considerably different from the reference and also rather implausible for the FI-source. In contrast, GMT60 was rated as perceptually identical and similarly plausible as the reference. A careful interpretation of these observations is that a scene may affect the estimated plausibility of the reference, if it sounds slightly different. This interpretation has to be looked at with caution, because also CRev50 and GMT40 were rated as perceptually identical in the similarity ratings and CRev as clearly different and significantly less plausible than the reference.

In addition, for the side-indirect sound source, again no significant effects of the test condition on the plausiblity of the reference could be observed.

An in-depth investigation is necessary to verify whether such effects can occur systematically.

Similarity - If both scenes in the test pair sound identical, the subjects had to rate similarity with *1*. However, even if both scenes actually were identical, the similarity was sometimes rated with *2*. That can happen, for example, if the participant slightly moves when switching between A and B or because the speech continuously progresses and the subject never has the chance to directly compare precisely the same sound.

For this reason, the similarity ratings are tested for significant differences from the similarity ratings of the *Orig-Orig* pair separately for each loudspeaker scenario. The anchor pair in the direct reproduction and *LR-CRev* in its function as the anchor pair in the indirect reproduction was not included in the analysis.

The paired, two-way Wilcoxon signed-rank tests with an mFDR-corrected significance level indicate that in the direct reproduction, *LR, CRev, GMT11* do not sound exactly like the reference for both loudspeaker positions. Their similarity to the original measurement was rated significantly different (all $p < 0.001$) from that of the reference. For the indirect reproduction scenario, only for *LR* and *CRev*, significant differences were found (both $p < 0.001$) for both source positions front and side. This is not surprising, because these modifications were also rated as significantly less plausible. For *LR50, CRev50, GMT40, and GMT60* in the indirect scenarios, no significant differences could be observed (all $p > 0.1$) for both source positions. These results suggest that all conditions with modifications starting around the predicted mixing-time sound identical to the original.

Free descriptions of the audible effects - Every time a participant did not rate the plausibility of a scene with "6 - clearly, Yes" or the similarity of a pair with "1 - same," the participant was asked to give reasons and explain their perception. This section summarizes the main issues.

Measured references - The original BRIR set was not consistently rated as "clearly plausible." One participant reported hearing the sound source slightly elevated in relation to the visible loudspeaker. For the "Side Direct" case, three participants reported perceiving slight inconsistencies in the reproduction in terms of a step-wise change of the localization, especially in the range in front of the side loudspeaker (Pos 2-4 of the translation line).

Direct reproduction scenario - For the various test cases in the direct scenario, plausibility was not rated significantly different from the original, except for the anchor. The anchor was described as

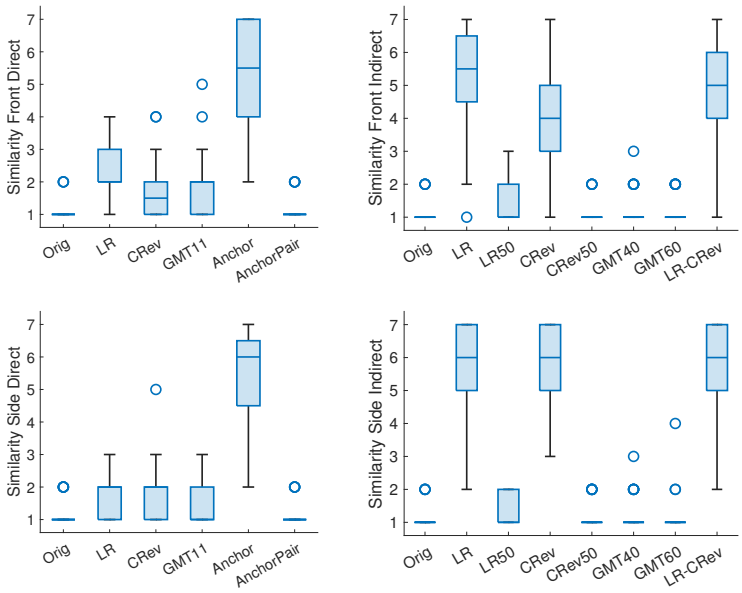


Figure 4.30.: Results for Similarity.

too dry, too small, and too close. The sound source was localized in front of the actual loudspeaker. It was also perceived as clearer and less blurry than the original. For *LR*, a tiny position shift to the right was reported. One participant estimated a shift of 1-2 cm. However, the auditory image was still in line with the visual impression. For *CRev* and *GMT11*, the majority of participants rated similarity with *2* because when switching between A and B, "something seemed to have changed", or they had "the feeling that it is different", without being able to describe it in more detail.

Indirect reproduction scenario - In the indirect scenario, *LR* and *CRev* were perceived less plausible for both source cases. For *LR(FI)*, participants reported perceiving surprisingly much reverberation coming from the back of the line. In some parts of the translation line, the localization shifted completely to the back. In *LR(SI)*, the source was reported to be on the left side, which was very incoherent with the visual impression of the loudspeaker. *CRev* was described similarly for FI and SI. The sound source appeared to move along during rotation and translation motion. Some participants also described that the sound seemed to come from all directions. For some positions and angles on the line, the *CRev* case seemed plausible. For this reason, many participants did not rate it as low as *LR*.
In contrast to the test cases where the reverberation was manipulated right after the direct sound, *CRev50*, *GMT40*, and *GMT60* were perceived as identical to the reference by most participants, for *LR* several participants reported minor audible differences in terms of coloration.

Overall, it may be summarized that for the modifications which seriously affected plausibility, the main reason was a wrong localization due to unnatural image shift effects caused by the impaired ER patterns.

4.5.4. Discussion and Conclusions

The presented study investigates the effects of BRIR modifications and unnatural spatiotemporal patterns of early reflections (ER) on the plausibility and similarity of auditory illusions created with position-dynamic binaural synthesis. Realistic AR scenarios were created by acoustic measurements in a seminar room. They included the cases of walking towards, past, and behind an ordinary loudspeaker which is representative of position-dynamic exploration of auditory scenes.

The results suggest that in cases of prominent direct sound (9-21 dB above the strongest ER), listeners exhibit low sensitivity to the physical details of the reverberation and the accuracy of the spatiotemporal ER patterns. For listening positions in front of the sound source, the simple constant-reverb approach is sufficient to create the plausible auditory illusion that the real loudspeaker reproduces speech. Even severe manipulations of the ER pattern, like a left-right swap, did not significantly decrease plausibility, although slight differences to the original data set were audible.

In contrast, in the indirect irradiation cases with low direct sound energy (0 dB to -14 dB relative to strongest ER), the same modification approaches caused severe degradation of plausibility and similarity. In particular, substantially wrong localization was stated by the participants as the reason for low plausibility ratings. If the BRIR impairment started at the predicted perceptual mixing time, 50 ms after the direct sound (or 40 or 60 ms second after the start of the sound emission), the scenes were perceived as identical to the reference with equal plausibility.

Generally, these observations suggest that considering a listening area in front of a directional sound source is not a critical test scenario to evaluate or even confirm the suitability of interpolation, extrapolation, or any other auralization approach. Scenarios with a low direct sound, like the indirect reproduction scenario considered in this study, will allow for more valid statements about the quality of an auralization realization for position-dynamic binaural synthesis. It is likely that several inter- or extrapolation approaches listed in the introduction would not perform well in the indirect scenario, for example, the ITDG-shaping approach proposed by Werner et al. [191], which does not correctly consider room geometry. Also, randomized ER patterns, as discussed by Arend et al. [85], can be difficult.

This case study confirmed that in the considered seminar room, a perceptual mixing time of 50 ms after the direct sound is sufficient. The results align with the models derived by Lindau et al. Lindau et al. [150] and with the conclusions drawn by Meesawat and Hammershøi Meesawat and Hammershøi [151]. Furthermore, as expected, no substantial perceptual differences could be observed for the two different approaches of applying the perceptual mixing time to position-dynamic auralizations, namely the Constant-Reverb and the Global Mixing Time.

However, for the indirect source in the front, the plausibility of the reference was rated significantly lower when compared to the two constant reverb-test cases. Future research will have to test for systematic impact.

This experiment considered only one room with four specific arrangements of the sound source, only a selected range of listening positions, only one sound source at a time in the room, and only male speech as a test stimulus. Further studies are necessary to draw more general conclusions. However, this study contributes to a better understanding of the relevance of a correct spatiotemporal ER pattern in position-dynamic binaural reproduction and its limits.

5. Discussion, Summary and Conclusions

This chapter first provides an in-depth discussion of the results and observations and reassesses them, considering other related studies. The goal is to examine which conclusion may be drawn overall. The second section of this chapter summarizes the research realized within this thesis and the third section lists the final conclusions of this work.

5.1. Discussion

As explained and discussed at the beginning in Sec. 1.7, the research conducted in this thesis investigated the following **research questions**:

- Is the implemented system capable of providing plausible illusions of walking towards, past, or away from the sound source?
- What is a suitable method to evaluate the plausibility of walking towards, past, or away from a virtual sound source?
- **Which psychoacoustic requirements can be defined for the accuracy of imitating the position-dependent room acoustic properties in a position-dynamic binaural auralization with interactive listener translation to create plausible auditory illusions of walking towards, past, and away from a sound source in a room?**

This section discusses which answers can be given based on the results of the investigations.

5.1.1. Validation of the realized system for position-dynamic binaural auralization of rooms

To study the perception of auditory illusions for listener translation, a system for position-dynamic binaural room auralization was realized. To be a valid tool for the desired investigations, the system needs to be capable of creating plausible spatial auditory illusions to be explored by the listener with self-translation and self-rotation (and head rotation). First, the system's technical limitations are listed, followed by a discussion of its capability of providing a plausible reference scene.

Technical limitations of the system

In experiments I-V, position-dynamic binaural auralization is realized based on BRIR data sets determined for discrete room positions in a uniform distribution along a line for listener translation or a listening area in the specific case of experiment III. The spatial resolution of the BRIR position grid and the angular resolution of the azimuth directions were limited. Listeners with varying head sizes, head shapes, body heights, and body shapes evaluated the system in terms of the plausibility of the created spatial auditory illusion. Neither of these individual variations was considered in the reproduction. No individualization was applied, and elevation changes were not considered.
Furthermore, a system-based pre-determined BRIR filter cannot consider motion-induced Doppler effects.

In experiments I, II, and III, STAX headphones were used together with an HTC Vive HMD. This combination was very stable on the head. Only the HMD limited the options for placing the headphones over the ears. This issue may have a slight impairment effect on headphone compensation because the actual transfer path slightly differed from the measured ones.

In experiments IV and V, the extra-aural AKG K1000 headphones were used without an HMD. However, these headphones have the disadvantage that they are relatively loose and unstable on the head. This issue might again slightly impair the headphone compensation, and in addition, the stability of the virtual sound source can be affected during quick motions of the head. Therefore, participants were instructed to consider this effect during their motion.

Resolution As one of the main limitations, the required minimum BRIR position grid resolution to achieve a perceptually continuous reproduction for changes of the listening position. Experiment III studied the continuity of the position-dynamic binaural reproduction with different resolutions. The results showed that for some signals, a uniform grid needs to have a resolution of 5 cm or even denser.

For example, Arend et al. [85] used a very similar auralization system based on a pre-determined BRIR set, but the BRIRs were time-aligned with respect to the direct sound. This can improve continuity and reduce the required minimum resolution for cases with strong direct sound. However, if this approach is applied to a scene with reflections of up to +10 or +14 dB, as in the indirect irradiation scenario in the seminar room, an alignment of the direct sound might lead to audible discontinuities. At least, in these cases, the behavior of auralizations without time-alignment will be better than a method based on time-aligned direct sound.

Werner et al. [260] proposed a non-uniform grid distribution wherein the grid points get denser with decreasing distance to the sound source. This approach can save many grid points and, thus, memory space carrying BRIR data. However, also this approach is only valid for cases with strong direct sound. In the indirect scenario, prominent early reflections may also have to be considered in the grid, for example, with additional dense areas.

According to the experimental results presented in section 4.3, the best resolution for speech was 5 cm. However, the two following experiments were still conducted with a distance of 25 cm between the measurement positions. This decision was taken for practical reasons, to keep the measurement effort reasonable. For distance changes, the results of experiment 4.1 indicated a satisfactory continuity which did not degrade plausibility. For the case of walking past the sound source with direct irradiation, few experienced listeners reported minor discontinuities, particularly when standing in front of the side loudspeaker, looking towards it, and turning their heads. In this case, their motion sometimes caused a switch to the neighbor position, although they were still standing at the same point. This was audible as a slight position shift of the virtual sound source. If there is a visible source object like the loudspeaker that remains stable, that may affect (audiovisual) plausibility. Some of the participants reporting this effect rated the plausibility lower. However, when walking at an ordinary speed, listeners usually did not notice discontinuities.

Perceptual effects caused by technical limitations of the test setup As a result of these technical limitations, the system is, for example, unsuitable for realizing position-dynamic auralizations for a source of noise signals. The results of the first experiment showed that if white noise is auralized with a positional resolution of 25 cm along the translation line, the noticeable degradation of continuity also affects the plausibility of walking towards the virtual sound source. With male speech, these effects appeared negligible. Only in a few cases some experienced listeners noticed slight imperfections in the continuity.

Furthermore, during scene exploration, the listener will usually not stay perfectly on the given line for listener translation. This might lead to slight instabilities in the localization of the sound source. Similar effects can occur due to the lack of individualization and the lack of considering the

differences in elevation due to deviations in body height and headphone misplacement. Particularly, slightly elevated localization of the virtual sound source was mentioned by some participants, which is very likely caused by the lack of individualization. Several people also reported that in the beginning, before moving, they experienced in-the-head-localization. However, with the originally measured BRIRs, after some movement, every subject confirmed good externalization.

Interestingly, despite all these limitations and drawbacks, the system was capable of providing plausible auditory illusions to most participants. The main goal of the developed setup was the realization of a plausible illusion of walking towards, past, and away from a virtual sound source. Furthermore, the goal was to realize a reference scene purely based on measured BRIRs without considerable interpolation or extrapolation to be able to measure degradation effects on plausibility caused by impairments or simplifications of the BRIR data set. The following paragraph discusses why it is reasonable to conclude that a virtual reference scene based on the 25 cm grid and male speech as a test signal can be regarded as a suitable, plausible reference.

Plausibility of the reference scenes

The plausibility of the position-dynamic binaural auralizations of the originally measured BRIR data sets was repeatedly evaluated in several experiments by different groups of participants and with different approaches to evaluation. Without the availability of the real counterpart of the simulated sound field, the vast majority of the participants rated the measured scenes as plausible. Even with the availability of the hidden real reference, many inexperienced listeners were not able to confidently detect the simulation. Experienced listeners were mostly able to distinguish the real version from the simulation once the real version was available. Before, when only the simulation could be heard, even the experienced candidates were not at all confident in identifying the simulation. One of the primary cues that revealed the virtual version in experiment IV could be traced back to the effects of headphones placements on the dummy head during the BRIR measurement and on the participant's head during the experiment. This marks the difficulties of appropriately simulating the shadowing effects of the headphones also in the virtual version to compare the reproduction to the real versions of the sound fields. This is rather a drawback of the test method than of the reproduction system.

5.1.2. Evaluating the plausibility of spatial auditory illusion for position-dynamic binaural auralization

In Sec. 1.5, plausibility was defined as the agreement with the own internal reference. It was also discussed in Sec. 2.7 that the individual internal references are subject to variations between different people as well as for the same person over time and in the context of listening. Generally, the correctness and sharpness of the own internal reference depend on the individual listening experience for the specific scenario. If a subject never listened to a similar sound field with the necessary attention, the internal reference may be wrong, and the expectation may differ from reality.

This thesis investigated the plausibility of spatial auditory illusions in various ways. Plausibility was evaluated in a Yes/No paradigm with and without considering the corresponding real versions of the simulated sound fields. It was rated on a 4-point interval scale in a single scene paradigm without a visual impression of the auralized acoustic scene and on an ordinal 6-point scale in a paired comparison with only virtual test items but with an additional visual impression of the auralized sound source and room environment, demanding for audiovisual coherence and thus adding a further dimension to the plausibility construct. This section discusses the suitability and limitations of the different test methods and the role of choosing a suitable reference.

Perceptual evaluation considering listener translation This thesis presents some of the early experiments considering interactive listener translation. The listener can control the avatar's movement with his/her own body motion. In the ideal case, the motion of both is equivalent and, thus, authentic. In everyday life, listeners have a certain expectation of how the sound in their ears should change according to their own motion.

One of the essential effects of dynamic binaural synthesis is that the listener can move relative to the sound objects. If a stationary sound source in a room is auralized, both should remain stable instead of moving along with the listener's motion. Furthermore, a continuous, natural progress of the dynamic sound adaptation is desired.

In order to evaluate all these aspects, the test design needs to be based on subjects interactively exploring the scene and walking to change their listening positions relative to the sound source and the room. This exploration phase usually takes longer than rating binaural auralization of specific stationary listening positions. Therefore, considerably fewer test cases can be investigated in an experiment conducted with volunteering participants. Furthermore, it is crucial that participants only rate the test case after a suitable amount of interactive exploration. Some encouragement was necessary with inexperienced listeners until they started developing their individual exploration behavior. As a consequence, a rule was set up for all the experiments: first, the participants had to walk up and down the 2 m translation line at least once before being allowed to rate the experience. Often, thereafter, listeners moving close to the sound source and exploring that place with relatively small movements was part of the exploration strategies. However, also relatively fast and prolonged position changes were used for exploration.

Different approaches of evaluating plausibility and the role of choosing a suitable reference Inspired by the method proposed by Lindau and Weinzierl [211], experiment IV in this thesis compared the simulated sound fields to their real counterparts. The real version was unavailable in the first part of the test, but the participants did not know that. It was interesting to notice that even experienced listeners could not confidently tell that they were only listening to the simulated sound field. Furthermore, the experiment showed that the availability of the real version among the test scenes increased the capability to identify the virtual version. It is not new that the availability and the choice of reference considerably impact the ratings in psychoacoustic studies. However, this experiment highlights the role of the reference regarding the plausibility of position-dynamic binaural auralizations.

Rather than comparing the virtual content to its real counterpart, Gospodarek et al. [276] proposed to evaluate the plausibility of a pair of a virtual and a real sound source placed at different positions in the room, reproducing sound consecutively. This is an interesting approach with regard to the ecological validity of common AR scenarios. Furthermore, they argue that rating plausibility on a continuous scale can give an insight into the correlation with other attributes like externalization, differences in timbre or reverberation, or localization accuracy. A similar approach was chosen in experiment I (Sec. 4.1) in this thesis, where continuity, externalization, and impression of walking towards a virtual sound source and in experiment II (Sec. 4.2) where sound source stability was rated in addition. The correlation analysis of the related psychoacoustic data confirms plausibility as a multidimensional construct for measuring the overall quality. Plausibility showed a relatively high Pearson correlation with each of the other constructs. In contrast, the correlation between the lower-order attributes, for example, externalization and continuity, was relatively low. Gospodarek et al. [276] announced a thorough analysis of the interdependencies. However, a publication thereof is still pending.

For plausibility, Gospodarek et al. [276] decided to use a scale from *0 - not at all plausible* to *6 - very plausible*. Interestingly, they report that the plausibility of real sound sources is also not always perceived as entirely plausible. This might be an indicator of an inaccurate internal reference which could lead to expectations (slightly) deviating from reality.

Schneiderwind and Neidhardt [328] conducted a similar study based on the *Yes/No* or *Real/Virtual*

paradigm with two loudspeakers present, which might both reproduce sound virtually or actually. Furthermore, Wirler et al. [217] proposed the concept of transfer plausibility and a corresponding test setup consisting of a varying amount of real sound sources to provide a reference environment rather than the real version of the virtual content.

Testing plausibility considering real versions of the sound fields still requires taking the effects caused by the presence of the headphones into account because perfect transparency has still not been achieved [277]. For experiment IV in the thesis, BRIRs were measured with headphones placed over the ears of the dummy head. The extra-aural headphones AKG K1000 were chosen due to their relatively high acoustic transparency [278]. However, realizations based on electro-acoustic hear-through solutions are an interesting approach to achieving transparency, too [279], [329]. For example, Doll [280] realized a comparison of a virtual and real sound source using an electro-acoustically transparent earpiece [281]. This pilot study indicated that a careful adjustment of the coloration is essential to avoid revealing the simulation.

An alternative to comparing realized virtual contents to their real counterparts are test designs based on a virtual reference. This approach was used in experiments I, II, and V within this thesis. However, this approach requires a plausible virtual reference it is worth comparing to. Determining the requirements for such reference and realizing it is one of the achievements of this Ph.D. thesis. This is discussed in more detail in the next section. For the current section, the availability of a suitable virtual reference is assumed.

Still, it has to be considered that virtual references are not necessarily perfect in their overall quality and their plausibility. Hence, the evaluation method has to take this fact into account. In the experiments I and II, a single scene/single source paradigm closely related to the Absolute Category Rating defined in ITU-T Rec. P.800 [271] was realized. Because multiple attributes were evaluated, the term *Multiple-Attributes Absolute Category Rating - MAACR* seems suitable. This approach has the advantage that a scene is evaluated after exploring and experiencing it actively. For this reason, the test method was proposed to the MPEG-I Audio committee in the ongoing standardization process, where suitable test methods for interactive, immersive audio for VR and AR are required. A disadvantage of MAACR are long test durations if several attributes are taken into account. Moreover, it requires that the participants have a certain level of expertise in understanding and evaluating the given attributes. In the experiment, most participants rated the measured references BRIR set as plausible, and most rated the scenes without reverberation as clearly not plausible. This was also the case in the repeated evaluation in experiment II.

To take the drawbacks of this method into account, experiment V evaluated plausibility based on a paired comparison. Throughout the scene exploration process, the participant could switch between the hidden virtual reference and the modified scene with the simplified reverberation and rate the plausibility for both realizations separately. This approach does not require that the reference is a perfect realization, but it still allows for a relative comparison between the reference and the test case of interest. Furthermore, it enables less experienced listeners to provide more critical ratings because the available reference will likely also have some tuning effect on the internal reference. The results of the experiment revealed substantial differences in plausibility, and the listeners repeated ratings showed relatively high consistency with this test method.

Tab. 5.1 attempts to show how the studies within this thesis cover the two categories of testing with regard to the pure internal reference and a tuned internal reference. However, these two categories are actually part of a continuum, which might even have more dimensions than one. Already experiment V allows for direct comparison to the virtual reference. Therefore, speaking of an evaluation with regard to the pure internal reference would not be correct. Moreover, the test design intends to not entirely rely on the individual internal references of the various test subjects. The attempt to add the other studies to the table, for example Gospodarek et al. [276] or Wirler et al. [217], is challenging because they would have to be placed somewhere between the categories.

Binaural Synthesis	Plausibility I pure internal reference	Plausibility II 'tuned' internal reference	Authenticity external reference
Static Reproduction	(✓)	Hartmann & Wittenberg 'Convincingness' (1996) Oberem et al. (2016) part B	Moore et al. (2010), Maseiro (2012), Oberem et al. (2016) part A
Head rotation	(✓)	Lindau et al. (2007), Lindau & Weinzierl (2012), Pike et al. (2014)	Brinkmann et al. (2017)
Rotation & Translation	**Initial Study (2017)** **Exp. I & II** **Exp IV - Part I**	**Exp IV - Part II**	

Table 5.1.: Summary of previous studies investigating plausibility and authenticity of binaural synthesis. Plausibility is split up into the two proposed categories of measuring the agreement with the pure internal reference or a tuned internal reference as a result of the indirect comparison with the real counterpart of the scene. This overview is not exhaustive but provides examples for each of the cases.

This simple framework requires more discussion and refinement in the future.

In summary, all the methods to evaluate plausibility realized within this thesis are suitable approaches, each with its own limitations, drawbacks, and advantages.

5.1.3. Position dependent room perception - required room accuracy in position-dynamic binaural audio

The investigations in the thesis considered a similar setup of a 2 m line for listener translation and two loudspeakers in two different rooms, the listening laboratory with $T_{30} = 0.23\,s$ and the seminar room with $T_{30} = 0.98\,s$. In both rooms, acoustic measurements were conducted with a Kemar HATS and a microphone array consisting of an omnidirectional microphone in the center with six satellite microphones in an open sphere arrangement with a 5 cm radius. The auralization was purely based on the measured BRIRs sets and systematic manipulations thereof. The microphone array data only served for detailed physical analysis.

Range of the room acoustical variations within a room Appendix A provides an overview of the physical changes along the translation line in both considered rooms. For example, the time-of-arrival, direction-of-arrival, and the energy relative to the direct sound are listed in Tab. A.2 for the listening laboratory and Tab. A.3 for the seminar room.

One major impact on position-dependent room perception is the position-dependent direct sound. Its level decreases over distance according to the inverse square law. Related ranges were discussed in experiment V.

The initial-time-delay-gap for the loudspeaker in front of the line ranged from about 4.5 ms at the closest position to about 2.3 ms at the farthest point of the translation line in both rooms.

However, the relevance of room acoustic parameters for describing and estimating local perceptual differences is limited.

Perception of variations in reflection patterns An obvious physical change with listening position (for a stationary sound source) can be observed in the temporal structure of the early reflection

pattern and changes in the direction of incidence.

Some own earlier work studied the perceptual differences at five different listening positions in a small conference room [130]. Furthermore, the capability to understand and assign an auralized perspective to the corresponding listening position was investigated considering two different training methods [323] [146].
We found that the deviations in the local acoustic properties caused audible differences in the apparent source direction, source distance, source width, and timbre. If the loudspeaker was turned away from the listener, also reverberance differed.
Although there are audible differences between the positions, most participants found it hard to assign the corresponding position based on an audio-only representation [188]. Additionally, the participants were provided with the corresponding visual perspectives in terms of a 360° image. The results showed that many participants could not localize the auditory perspective in the room.

Similar observations were made by Meyer-Kahlen and Schlecht [282], Meyer-Kahlen et al. [283]. However, solving this assignment task is a matter of listening experience and training. It is known from many studies that the evaluation of minor room acoustic differences can often only be perceived or understood by trained room acousticians.

If listeners who hear differences cannot interpret or understand what they hear, this suggests that random changes in the early reflection patterns might be an easy-to-generate alternative to the actual pattern that is usually hard to estimate.

Role of listener expertise for room acoustics Evaluating the perception of rather small deviations in room acoustics requires experienced listeners. Therefore, it is desirable to invite test participants with room acoustic listening expertise for the related experiment. Often, experimenters distinguish experienced and naive listeners but do not specify the expertise and how it was assessed. Commonly, a general interest in music or musical engagement is assumed to be a good predictor for room acoustic listening expertise.

von Berg et al. [284, p. 1] constructed a series of experiments to investigate the "perceiving and understanding of room acoustical phenomena" and examine what constitutes acoustic expertise in general and which indicators could help measure room acoustic listening expertise. They conclude that the traditional criteria for choosing expert listeners are limited in their capability to predict their actual performance in the perceptual assessment of room acoustics.

In all five experiments conducted within this thesis, similar effects were observed. Even staff members in the related field did not necessarily perform well. Also, the ability to play an instrument did not necessarily lead to a critical assessment. Furthermore, it was evident that inexperienced listeners are only of limited help in assessing minor deviations. However, even inexperienced subjects noticed if the auditory perception was not in line with their own movements. Since humans are used to listening during motion in their everyday lives, maybe the internal reference of average listeners is more reliable with respect to motion than regarding room acoustic properties.

Main observations on the perceptual effects of simplified room representations

In this thesis, it has been shown repeatedly that keeping the reverberation independent of the position for the tested 2 m line does not affect the plausibility of the interactive approaching motion in front of the loudspeaker. This is true for both rooms. However, in the reverberant seminar room, it made a difference whether the BRIR position to extrapolate from was taken from inside the critical distance or outside. However, in a direct comparison, audible differences can occur.

When the original BRIR data for walking towards a sound source is auralized with noise, each step closer to the sound source yields a different apparent pitch, probably caused by the comb filter due to the floor/ceiling reflections. The apparent pitch increases with decreasing distance to the source. It underlined the discrete steps covered by the filters and affected the perceived continuity.

Furthermore, it is also an example highlighting the relevance of accurately rendering the prominent early reflections rather than keeping the whole reverberation constant.

In experiment V, the indirect irradiation scenario challenged the constant-reverberation case and revealed its limitations. Prominent localization errors occurred and basically destroyed any plausibility.

Relevance of the image shift effect for plausibility and related room simplification The observations of Experiment V suggest that the image shift effect is highly relevant for audiovisual plausibility. Changes in the image shift effect have to be avoided during binaural rendering. Therefore it would be interesting to have a model that indicates whether early reflections substantially contribute to the current localization of the auditory source image. If reflections contribute, they must be rendered accurately to preserve the localization.

[285] was concerned with predicting the occurrence of image shift effects in concert halls with the goal of avoiding them. The audience would like to localize the musicians where they see them instead of somewhere else. He proposed a prediction based on the early interaural cross-correlation coefficient (1-IACC$_E$) in the octave bands around 500 Hz, 1 kHz, and 2 kHz. According to Okano's analysis, the frequency range around 250 Hz and below usually does not contribute to the image shift effect.

The effects of damaged localization mainly occurred in the indirect irradiation scenarios. Tab. A.3 provides an overview of the properties of the six most prominent reflections at both ends of the translation line in the seminar room. In the indirect cases, these reflections exhibit an increased level compared to the direct sound. According to Barron's visualization [115] (Sec. 2.4.1), at these levels, the image shift effect can occur at a wide range of delay times. Similar observations were made by Pastore and Braasch [286], who investigated cases with increased lag levels in the lead-lag paradigm. Related perceptual models are still pending but could be helpful for reasonably simplifying room representations and psychoacoustic optimizations of binaural rendering.

Late reverberation and perceptual Mixing Time Several studies indicate that also late reverberation contains audible directional differences [77, 158, 149]. Furthermore, Lindau et al. [150] explain that extending the concept of the perceptual mixing time might be challenging due to potential impacts of room modes. These would lead to position-dependent audible differences. Experiment V showed that even for the indirect irradiation case, the perceptual mixing time predicted according to their signal-based and system-based models were suitable predictions. The modifications in the reverberation after the mixing time mostly do not even cause any audible differences. These observations confirm the validity of the mixing time predictions and, in addition, suggest that the concept is also interesting for position-dynamic auralization.

The Schroeder frequency for the studied seminar room is about 165 Hz, and the auralization covered a range down to about 60 Hz. However, any perceptual effects of room modes did not play a role in the evaluation. None of the participants made any statement about room modes. However, the translation line and the loudspeaker were set up in the room in a slightly diagonal arrangement. Hence, the setup did not particularly facilitate the occurrence of audible room modes.

However, for position-dynamic auralizations of small rooms, the role of room modes should be kept in mind when simplifying the late reverberation by a constant transfer function.

Simplification based on audibility threshold Hacıhabiboğlu and Murtagh [132] and Brinkmann et al. [80] proposed the idea that reflections underneath the audibility threshold do not have to be rendered. Both realized related investigations with room simulations based on the mirror image source model. Brinkmann et al. claim that in most cases, only about six audible early reflections would have to be rendered separately.

With the tools provided by Brinkmann et al. [80], the SRIRs measured with the microphone array were analyzed. The resulting visualizations can be seen in fig. A.5. The reflections determined as audible by the implemented model are marked in red. The number of salient reflections for Pos1 in front of the loudspeaker is around six. However, already at Pos9, the model considers many more reflections audible. The third graph visualizes the situation for the indirect irradiation by the front loudspeaker for a listener located at Pos9. For this case, a simplification based on the audibility threshold would not significantly simplify the auralization. However, it should be kept in mind that cases like this were not considered in developing models for the audibility of room reflections. It is likely that also the indirect scenario bears lots of potential for perception-based simplification. Only related models need to be extended to cover such cases.

5.1.4. Remaining Open Questions - Future work

First, it seems interesting to study the indirect reproduction scenario in the listening laboratory, considering that their level is constrained by the ITU-R BS.1116. Overall, the results of the experiments conducted with the two selected test rooms do not allow drawing conclusions for any other room. However, it seems reasonable to expect comparable effects for rooms with very similar acoustic conditions. More studies considering different room properties, degrees of reverberance, and specific spatiotemporal patterns of early reflections are important. It seems of particular interest to better determine the audibility threshold and understand the various parameters contributing to the occurrence of audible effects caused by specific room reflections in an ensemble. Especially the last experiment also highlighted the role of the image shift effect in audiovisual plausibility. Consequently, it seems relevant to further explore this apparent shift of the source location and determined related parameters and thresholds. Developing a model of the image shift effect and the precedence effect can help increase the efficiency of position-dynamic binaural rendering. Furthermore, exploring simplified early reflection representations in the indirect irradiation scenario is of interest.

The perceptual effects of single reflections and the overall room response are closely related to the type of signal. Therefore, it would also be important to conduct similar experiments with different signals like white or pink noise. Noise is a more critical test signal and reveals audibility differences more easily. In this thesis, this was shown especially for the similarity to the measured reference, but also in terms of continuity and consequently plausibility. Conducting similar experiments with noise will require a different implementation with continuous changes between different listening positions.

5.2. Summary

This thesis investigates the influence of simplified acoustic room representations on the plausibility of dynamic binaural auralizations considering listener translation. The role of physical inaccuracies in the spatiotemporal pattern of early reflections is one of the key questions. It is of particular interest for efficient audio realizations for VR/AR/MR/XR because an accurate imitation of the early room reflections can be very costly in terms of computation power and memory.

Before starting the investigations for this thesis, a system for position-dynamic binaural auralization had to be implemented. On the one hand, the flexible Python tool pyBinSim was developed based on previous work by Frank Wefers [62] to realize partitioned convolution for binaural auralizations in real-time and select the filters of a BRIR data set according to the listener's motion. On the other hand, BRIR data sets for a 2 m long line for listener translation based on measurements with a KEMAR 45ba head-and-torso-simulator and room acoustic simulations were created. With the given system, listeners could walk towards, past, and away from a virtual sound source. With the originally measured BRIRs sets and male speech as the source signal, a plausible illusion of walking

towards and away from the sound source could be created for most participants. However, with white noise, the continuity decreases substantially, negatively impacting plausibility.

Consequently, the required minimum resolution for uniform BRIR position grids was examined for different source signals. For speech and white noise, a grid resolution of 5 cm achieved the best-perceived continuity. For a solo saxophone piece rich in transients, even a 5 cm resolution did not yield optimal continuity. This suggests that a continuous parametric adaptation of the reproduction to the listening position is desirable. However, these values were estimated in a procedure where listeners moved very cautiously and paid lots of attention to finding discontinuities. Likely, a slightly lower resolution in average listening scenarios may also be sufficient. Continuity was also evaluated for the auralizations of measured BRIR data sets to ensure sufficient continuity in the experiments on room perception during listener translation. The results indicate that good continuity was achieved.

Furthermore, the measured BRIR sets are supposed to serve as references in the psychoacoustic investigations. Therefore, the plausibility of the original auralizations was evaluated with different test methods and varying contexts. Besides a single scene evaluation that might be best described as Multi-Attribute-Absolute-Category-Rating (MAACR), an evaluation based on the Yes/No paradigm proposed by Lindau and Weinzierl [211] was conducted with and without a real counterpart as hidden reference. Without the availability of a real reference, inexperienced listeners mostly accepted the auditory illusion as real, and experienced listeners could not confidently identify the simulation. When the real version of the sound fields was added to the test scenes, experienced listeners were mostly able to recognize the virtual sound source and inexperienced listeners at least started to doubt the realness of the imitation. However, it has to be kept in mind that the consideration of real sound fields in the evaluation of headphone-based auralizations requires taking the effect of the headphones placed over the listeners' ears into account. The participants' comments suggest that the non-individual consideration of the headphones was one of the primary cues revealing the simulation. In addition, slight instabilities during fast head rotation were stated as revealing cues. This might be due to the use of generic BRIRs, instability of the headphones on the head, or insufficient system latency.

In the final study, the plausibility of the originally measured scenes and the manipulated scenes was evaluated in a paired comparison. This seems a reasonable compromise that does not require taking the shadowing effects of headphones into account but provides an option to compare to a virtual reference while evaluating. The reference may not be assumed to create perfect plausibility for every system user. This approach evaluates differences in plausibility to the reference. However, the reference achieved results in a very high range of the scale. Experiment IV shows that the real version of the sound does not necessarily achieve perfect plausibility ratings. Only a direct comparison to the real counterpart considering the headphones' influence, can evaluate the top end of perceptual realism. The evaluation of plausibility will always be limited by the correctness and accuracy of each listener's internal reference.

In three experiments (I, II, and V), systematic modifications were applied to the measured BRIR data sets to study the main research question on the perceptual consequences of simplified or impaired room representations. Experiments I and II explored the boundaries of the plausibility of walking towards/away from a sound source in the two rooms by quite severe BRIR set modifications by removing the position-dependency of pre-delay, the initial time delay gap (ITDG), the direct-to-reverberant-energy-ratio (DRR) as well as the Spatio-temporal pattern of early reflections (ER). The results suggest an essential role of the DRR and its change in both rooms, but for the tested line and male speech as the test signal, the natural change of pre-delay, ITDG, and the ER patterns were not substantial for plausibility in front of the loudspeaker in the MAACR test design. In both rooms, a plausible impression of walking towards a virtual sound source could be achieved by using the BRIRs measured at one position of the 2 m-line and only adjusting the direct sound level according to the distance change. However, in the reverberant seminar room, choosing a position outside the critical distance was advantageous for plausibility. This position-independent reverberation approach is very interesting regarding efficiency since it saves much computational effort compared to many other

suggested approaches. However, it is essential to understand its limitations.

Experiments III and V in this thesis aimed to investigate the perceptual consequences of unnatural ER patterns to explore the boundaries of position-independent reverberation. The modifications were evaluated in the audiovisual AR scenario in a paired comparison with the measured reference scene. Besides walking towards/away from the sound source, cases of walking past and behind the source were included. For listening positions behind the sound source, the most prominent early reflections achieved levels of up to nearly +14 dB relative to the direct sound. Modifying spatiotemporal patterns for reflections with a level above the direct sound could lead to substantial changes in the apparent source position. Such severe image shift effects seriously affected audiovisual plausibility.

Further outcomes In addition to the new insights into the perception of position-dependent room auralization, the thesis provides other valuable outcomes to the scientific community. On the one hand, the developed Python tool *pyBinSim* has evolved. It now provides options for further configuration, importing acoustic impulse responses in a more convenient format and importing them in different sections, for example, for separate handling of the early and late parts of the filters. Meanwhile, pyBinSim has been used for new experiments at various other institutes.

On the other hand, a different data set of acoustic impulse responses measured with the KEMAR 45ba HATS, as well as the microphone array data captured to analyze the early reflection pattern, was created in the course of this thesis. It is documented in detail [287] and is publicly available online[1]. An in-depth analysis of the physical properties is presented in appendix A. The data set is interesting in studying the plausibility of position-dynamic binaural audio in general and developing algorithms for auralization based on interpolation, extrapolation, and parameterization and conducting a physical evaluation in comparison to the measured data.

5.3. Conclusions

This thesis focuses on acoustic information measured in an actual room, taking into account all the acoustic details caused by the complex geometry of the seminar room with ordinary furniture as well as non-idealized sound sources, a two-way loudspeaker.

Based on the studies conducted within this thesis, the following conclusions can be made:

- The setup developed to realize position-dynamic binaural auralizations with a BRIR data set measured in intervals of 25 cm along a line for listener translation is capable of providing plausible auditory illusions of walking towards, past, and away from a virtual sound source. This was only shown for male speech as the source signal. In contrast, for example, with white noise, the plausibility was substantially affected by noticeable limitations of the continuity.

- In this thesis, spatial auditory illusions were created based on measured and simulated BRIRs without individualization. Repeatedly, most participants rated the measured reference, as well as several simplified scenes, as plausible. This means achieving plausibility does not necessarily require individual head-related cues. This observation is in line with the findings by Pike [288]. However, considering individualization might still improve the quality of auditory illusions.

- The required minimum BRIR position grid resolution was shown to depend on the signal and its frequency content. For white noise, a uniform grid with a resolution of 10 cm or worse, continuity was affected compared to a 5 cm resolution. For signals with strong transients like the solo saxophone excerpt, even the resolution 5 cm was not sufficient to provide a convincing continuity of the imitated sound field if the reproduction is not time aligned over the positions as, for example, in [85]. The test case in front of the loudspeaker in the quite dry listening

[1] https://zenodo.org/record/3457782 and https://zenodo.org/record/7838178

laboratory was chosen to examine the perceived continuity in dependence of the BRIR grid resolution. It puts the spotlight on the continuity of the dominant direct sound. However, this allows for drawing conclusions also for cases of prominent specular room reflections, which are unlikely to be more dominant in any room than the direct sound in the tested one.

- The availability of a real (hidden) reference can tune the listener's internal reference and affect not only the plausibility estimates but the actual acceptance of a spatial auditory illusion as real sound objects. This also suggests an important impact of the real acoustic environment on the plausibility of auditory illusion in AR/MR/XR scenarios.

- In a room meeting the ITU-R BS.1116, within the critical distance of the sound source, a plausible impression of walking towards and away from a sound source can be achieved by only adjusting the level of the direct sound according to the distance change with respect to the sound source. Keeping the reverberation with its spatio-temporal ER pattern independent of the listening position did not cause a measurable influence on plausibility. A constant transfer function representing the reverberation of the room does not significantly affect the plausibility. Not even a slightly randomized variation of the room response is required.

- In a reverberant room similar to the tested seminar room, it is also possible to create a plausible impression of walking towards and away from the loudspeaker by adjusting only the energy of the direct sound but keeping the reverberation independent of the listening position. However, in this room it was advantageous to choose a position outside the critical distance for measuring the BRIRs to extrapolate from. Further studies should investigate the generalizability of this observation.

- For a critical perceptual evaluation of algorithms for position-dynamic binaural auralizations based on interpolation, extrapolation, and parameterization of measured or simulated acoustic room information, the level of the direct sound has to be considered. Test scenarios with a listener in rather advantageous positions in front of the sound source are not sufficient. In particular for the imitation of directional sound sources in rooms, it is important to design test scenarios considering cases with low direct sound energy. In these cases, the geometric and acoustic details of the room reflections and their progress over the changing listening position are of higher perceptual relevance.

- Specular room reflections that arrive within about 8 ms after the direct sound or that exhibit an energy level similar to that of the direct sound or higher are likely to cause image shift effects which results in an apparent change of the source position. In these cases, accurate modeling of the geometric and acoustic details of the spatio-temporal pattern of early reflection and their progress over different listening positions is essential for the plausibility of the spatial auditory illusion.

- Current attempts to model the audibility threshold and consider it for binaural rendering requires further investigation. A state-of-the-art model indicates that most reflections may be audible behind a common directional sound source. However, it remains open whether, in such cases, most reflections contribute to the listening experience or whether the models need further refinement.

- Plausible auditory illusions can still exhibit substantial audible differences from the real reference. However, this thesis has shown that room representations with considerable simplifications can be perceived as very similar or identical compared to simulation based on a fully measured BRIR dataset. This suggests that a realization of authentic room auralizations can still be based on simplified physics, for example, considering the perceptual mixing time.

This thesis puts a spotlight on position-dynamic auralizations based on a constant transfer function representing the reverberation independent of the position and direction of the source and listener. This does not imply a recommendation for this auralization approach. Instead, this example is an interesting further reference, not for the highest possible quality, but for achieving very high quality at meager costs. Any more complex rendering should only be considered if it achieves significantly better perceptual quality for the desired scenario. However, particularly the indirect irradiation cases in experiment V revealed some of the limitations of the realization based on constant reverberation. With test signals that facilitate the audibility of comb filter effects, for example, white noise, a lack of position-dependent changes in the coloration might also affect plausibility for experienced listeners. Therefore, it is reasonable to explore improved approaches that consider position-dependent acoustic properties. For example, the mixing-time concept has also proved its validity in the test scenarios considered in this thesis.

Bibliography

[1] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[2] P. Zahorik, D. Brungart, and A. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica 91(3), pp. 409-420, May*, 2005.

[3] R. Nicol, *Binaural Technology*, ser. AES Monograph. Audio Engineering Society Inc., New York, 2010.

[4] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display*. J. Ross Publishing, 2013.

[5] DIN EN ISO 9613-2, "Acoustics - attenuation of sound during propagation outdoors - part 2: General method of calculation (iso 9613-2:1996)," 1999.

[6] M. Kleiner and J. Tichy, *Acoustics of Small Rooms*. CRC Press, Taylor and Francis Ltd., 2014.

[7] C.-H. Jeong, "Diffuse sound field: challenges and misconceptions," in *Inter-Noise, Hamburg, Germany*, 2016.

[8] M. Long, *Architectural Acoustics*, 2nd ed. Academic Press, 2014.

[9] S. Int. Organization for Standardization, Geneva, "DIN EN ISO 3382-2. Acoustics - measurement of room acoustic parameters - part 2: Reverberation times in ordinary rooms," 2008.

[10] H. Kuttruff, *Room Acoustics*, 6th ed. CRC Press, Taylor & Francis Group, 2017.

[11] M. R. Schroeder and K. H. Kuttruff, "On frequency response curves in rooms. Comparison of experimental, theoretical, and Monte Carlo results for the average frequency spacing between maxima," *J. Acoust. Soc. Am.*, vol. 34, no. 1, pp. 76–80, 1962.

[12] S. Olive and E. Floyd, "Detection of reflections in typical rooms," *J. Audio Eng. Soc., Vol. 37, No. 7/8, July/August*, 1989.

[13] U. Proske and S. C. Gandevia, "The proprioceptive senses: their roles in signaling body shape, body position and movement, and muscle force," *Physiological Reviews*, 2012.

[14] W. R. Thurlow, J. W. Mangels, and P. S. Runge, "Head movements during sound localization," *J. Acoust. Soc. Am.*, vol. 42, no. 2, pp. 489–493, 1967.

[15] P. Mackensen, "Auditive localization. head movements, an additional cue in localization," Ph.D. dissertation, University Of Technology Berlin, Germany, 2004.

[16] A. Honda, H. Shibata, S. Hidaka, J. Gyoba, Y. Iwaya, and Y. Suzuki, "Effects of head movement and proprioceptive feedback in training of sound localization," *i-Perception, 4(4), pp. 253–264*, 2013.

[17] H. Pöntynen and N. H. Salminen, "Resolving front-back ambiguity with head rotation: The role of level dynamics," *Hearing Research*, vol. 377, pp. 196–207, 2019.

[18] C. Kim, R. Mason, and T. Brookes, "Head movements made by listeners in experimental and real-life listening activities," *J. Audio Eng. Soc., Vol. 61, No. 6*, 2013.

[19] O. Brimijoin, W, A. W. Boyd, and M. A. Akeroyd, "The contribution of head movement to the externalization and internalization of sounds," *PloS ONE*, vol. 8, no. 12, p. e83068, Dec. 2013.

[20] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. Katz, and C. De Boishéraud,

"Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 2011–2023, Mar. 2017.

[21] H. M. Kondo, D. Pressnitzer, I. Toshima, and M. Kashino, "Effects of self-motion on auditory scene analysis," *Proceedings of the National Academy of Sciences*, vol. 109, no. 17, pp. 6775–6780, 2012.

[22] L. Wallmeier and L. Wiegrebe, "Self-motion facilitates echo-acoustic orientation in humans," *Royal Society Open Science, 1: 40185*, 2014. [Online]. Available: https://doi.org/10.1098/rsos.140185

[23] B. K. Shaw, R. S. McGowan, and M. T. Turvey, "An acoustic variable specifying time-to-contact," *Ecological Psychology*, vol. 3, no. 3, pp. 253–261, 1991.

[24] R. Guski, "Acoustic tau: An easy analogue to visual tau?" *Ecological Psychology*, vol. 4, no. 3, pp. 189–197, 1992.

[25] H. Wu, S. W. Y. Lee, and H. Y. Chang, "Current status, opportunities and challenges of augmented reality in education," *Computers & Education, 62: 41-49.*, 2013.

[26] P. Novo, "Auditory Virtual Environments," in *Communication Acoustics*. Blauert, Jens, 2005, ch. 11.

[27] P. Milgram, H. Takemura, and A. Utsumi, "Augmented reality: A class of displays on the reality-virtuality continuum," in *Proceedings of the SPIE: Telemanipulator and Telepresence Technologies, 2351: pp. 282–292*, 1995.

[28] R. Skarbez, M. Smith, and M. C. Whitton, "Revisiting Milgram and Kishino's Reality-Virtuality Continuum," *Front. Virtual Real., 24 March , Sec. Virtual Reality and Human Behaviour*, 2021. [Online]. Available: https://doi.org/10.3389/frvir.2021.647997

[29] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the authenticity of individual dynamic binaural synthesis," *J. Acoust. Soc. Am., Vol. 142, No. 4, pp. 1784–1795*, Oct. 2017. [Online]. Available: http://dx.doi.org/10.1121/1.5005606

[30] C. Kuhn-Rahloff, "Prozesse der Plausibilitätsbeurteilung am Beispiel ausgewählter elektroakustischer Wiedergabesituationen," Ph.D. dissertation, TU Berlin, 2011.

[31] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory for virtual acoustic environments (SAQI)," in *EAA Joint Symp. on Auralization and Ambisonics, Berlin*, 2014.

[32] C. Pike and H. Stenzel, "Direct and indirect listening test methods - a discussion based on audio-visual spatial coherence experiments," in *143rd AES Convention, paper 9829, New York, NY, USA*, 2017.

[33] B. F. G. Katz and R. Nicol, "Binaural Spatial Reproduction (Ch. 11)," in *Sensory evaluation of sound*, N. Zacharov, Ed. CRC Press, Taylor & Francis Group, 2019, pp. 349–386.

[34] J. Yang, A. Barde, and M. Billinghorst, "Audio Augmented Reality: A Systematic Review of Technologies, Applications, and Future Research Directions," *J. Audio Eng. Soc., Vol. 70, No. 10, pp. 788-809*, 2022.

[35] M. Vorländer, D. Schröder, S. Pelzer, and F. Wefers, "Virtual reality for architectual acoustics," *Journal of Building Performance Simulation*, 2015.

[36] B. Seeber, "What can we learn from simulated acoustic environments?" *Proceedings of Meetings on Acoustics, Vol. 19, 050196*, 2013.

[37] G. C. Stecker, "Using Virtual Reality to assess auditory performance," *The Hearing Journal, 72(6), pp 20–23, June*, 2019.

[38] H. Y. Seol, S. Kang, J. Lim, S. H. Hong, and I. J. Moon, "Feasibility of virtual reality audiological testing: Prospective study," *JMIR Serious Games, 9(3):e26976*, 2021.

[39] A. Ahrens, K. Duemose Lund, M. Marschall, and T. Dau, "Sound source localization with varying amount of visual information in virtual reality," *PloS one, 14(3), e0214603*, 2019.

[40] F. Pausch, L. Aspöck, M. Vorländer, and J. Fels, "An extended binaural real-time auralization system with an interface to research hearing aids for experiments on subjects with hearing loss," *Trends in Hearing, Vol. 22, pp. 1-32*, 2018. [Online]. Available: https://doi.org/10.1177/2331216518800871

[41] F. Pausch and J. Fels, "Localization performance in a binaural real-time auralization system extended to research hearing aids," *Trends in Hearing 24(10), pp. 1-18*, 2020. [Online]. Available: https://doi.org/10.1177/2331216520908704

[42] M. M. E. Hendrikse, G. Llorach, V. Hohmann, and G. Grimm, "Movement and gaze behavior in virtual audiovisual listening environments resembling everyday life," *Trends in Hearing, Vol. 23: 1–29*, 2019.

[43] G. Parseihian and B. Katz, "Rapid head-related transfer function adaption using a virtual auditory environment," *J. Acoust. Soc. Am. 31(4), p. 2948, April*, 2012. [Online]. Available: https://doi.org/10.1121/1.3687448

[44] H. Wierstorf, A. Raake, and S. Spors, "Assessing localization accuracy in sound field synthesis," *J. Acoust. Soc. Am.*, vol. 141, no. 2, pp. 1111–1118, 2017.

[45] V. Erbes, S. Weinzierl, and S. Spors, "Analysis of a spatially discrete sound field synthesis array in a reflective environment," in *EuroNoise Conference, Maastrich, Belgium*, 2015.

[46] V. Erbes and S. Spors, "Influence of the listening room on spectral properties of wave field synthesis," in *Annual Conference on Acoustic, Kiel, Germany*, 2017.

[47] M. Vorländer, *Auralization*, 1st ed. Springer-Verlag Berlin Heidelberg, 2008, doi 10.1007/978-3-540-48830-9.

[48] P. Zahorik, "Auditory display of sound source distance," in *Int. Conf. on Auditory Display, Kyoto, Japan*, 2002.

[49] C. Medonça, P. Mandelli, and V. Pulkki, "Modeling the Perception of Audiovisual Distance: Bayesian Causal Inference and Other Models," *PLoS ONE, 11(12): e0165391*, 2016, doi:10.1371/journal.pone.0165391.

[50] S. Nielsen, "Auditory distance perception in different rooms," *J. Audio Eng. Soc., Vol. 41, No. 10, pp. 755-770*, 1993.

[51] S. Int. Organization for Standardization, Geneva, "DIN EN ISO 3382-1. Acoustics - measurement of room acoustic parameters - part 1: Performance spaces," 2009.

[52] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in *108th AES Convention, paper 5093, Paris, France*, 2000.

[53] Z. Schärer and A. Lindau, "Evaluation of equalization methods for binaural signals," in *126th AES Convention, paper 7721, Munich, Germany*, 2009.

[54] A. Lindau and F. Brinkmann, "Perceptual Evaluation of the Headphone Compensation in Binaural Synthesis Based on Non-individual recordings," *J. Audio Eng. Soc., Vol. 60 (1/2), pp. 54-62*, 2012.

[55] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Binaural Technique: Do we need individual recordings," *J. Audio Eng. Soc., Vol. 44, pp. 451-469*, 1996.

[56] P. Minnaar, J. Plogsties, S. K. Olesen, F. Christensen, H. . Møller, and M. F. Sørensen, "The interaural time differences in binaural synthesis," in *108th AES Convention, Paris, France*, 2000.

[57] H. Møller, D. Hammershøi, C. B. Johnson, and M. F. Sørensen, "Evaluation of artificial heads in listening tests," *J. Audio Eng. Soc., Vol. 47, pp.83-100*, 1999.

[58] C. Mendonça, "A review on auditory space adaptations to altered head-related cues," *Front Neuroscience*, vol. 8, no. 219, 2014.

[59] C. Pörschmann, "3-d audio in mobile communication devices: Methods for mobile head-tracking," *Journal of Virtual Reality and Broadcasting, Volume 4, No. 13*, 2007.

[60] A. Lindau, "The perception of system latency of dynamic binaural synthesis," in *NAG/DAGA Int. Conf. on Acoustics, Rotterdam, The Netherlands*, 2009.

[61] M. Wenzel, Elizabeth, "The impact of system latency on dynamic performance in virtual acoustic environments," in *15th Int. Congress on Acoustics and 135th Meeting of the Acoustical Society of America*, 1998.

[62] F. Wefers, "Partitioned convolution algorithms for real-time auralization," Ph.D. dissertation, RWTH Aachen, 2014.

[63] C. Pörschmann, P. Stade, and J. M. Arend, "Binauralization of Omnidirectional Room Impulse Responses - Algorithm and Technical Evaluation," in *20th Int. Conf. on Digital Audio Effects (DAFX), Edinburgh, Scotland*, 2017.

[64] U. Sloma, F. Klein, S. Werner, and T. Pappachan Kannookadan, "Synthesis of binaural room impulse responses for different listening positions considering the source directivity," in *147th AES Convention, New York, NY, USA, 2019*, Oct 2019.

[65] V. Garcia-Gomez and J. J. Lopez, "Binaural room impulse responses interpolation for multimedia real-time applications," in *144th AES Convention, Milan, Italy, 2018*, May 2018.

[66] S. Garí, W. Brimijoin, H. Hassanger, and P. Robinson, "Flexible binaural resynthesis of room impulse responses for augmented reality research," in *EAA Conference on Spatial Audio Signal Processing, Paris, France*, 2019.

[67] P. Stade, "Perzeptiv motivierte, parametrische Synthese binauraler Raumimpulsantworten," Ph.D. dissertation, Technische Universität Berlin, Germany, 2018.

[68] M. Zaunschirm, M. Frank, and F. Zotter, "Binaural rendering with measured room responses: First-order ambisonic microphone vs. dummy head," *Applied Sciences*, vol. 10, no. 5, p. 1631, 2020.

[69] K. Müller and F. Zotter, "Auralization based on multi-perspective ambisonic room impulse responses," *Acta Acustica*, vol. 4, no. 6, p. 25, 2020.

[70] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger, and M. Marschall, "Higher-order spatial impulse response rendering: Investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution," *J. Audio Eng. Soc.*, vol. 68, no. 5, pp. 338–354, 2020.

[71] I. Engel and L. Picinali, "Reverberation and its Binaural Reproduction: The Trade-off between Computational Efficiency and Perceived Quality," in *Advances in Fundamental and Applied Research on Spatial Audio*, 2022.

[72] H. Kim, L. Remaggi, P. J. Jackson, F. M. Fazi, and A. Hilton, "3D room geometry reconstruction using audio-visual sensors," in *Int. Conf. on 3D Vision (3DV)*, Qingdao, China, 2017, 2017, pp. 621–629.

[73] T. Kim, Y. Kwon, and S. E. Yoon, "Real-time 3-D mapping with estimating acoustic materials," in *IEEE/SICE International Symposium on System Integration (SII). Honolulu, HI, USA, pp. 646–651*, 2020.

[74] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *J. Acoust. Soc. Am.*, vol. 138, no. 2, pp. 708–730, Aug 2015.

[75] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2746–2760, 2019.

[76] S. Bilbao and B. Hamilton, "Wave-based room acoustics simulation: Explicit/implicit finite volume modeling of viscothermal losses and frequency dependent boundaries," *J. Acoust. Soc. Am.*, 2017.

[77] B. Alary, A. Politis, S. Schlecht, and V. Välimäki, "Directional feedback delay network," *J. Audio Eng. Soc.*, vol. 67, no. 10, pp. 752–762, 2019.

[78] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, 2012.

[79] V. Bruschi, S. Nobili, S. Cecchi, and F. Piazza, "An innovative method for binaural room impulse responses interpolation," in *148th AES Convention, Online, 2020*, May 2020.

[80] F. Brinkmann, H. Gamper, N. Raghuvanshi, and I. Tashev, "Towards encoding perceptually salient early reflections for parametric spatial audio rendering," in *148th AES Convention, paper 10380, Online*, May 2020.

[81] A. Neidhardt, A. Ignatious-Tommy, and A. D. Pereppadan, "Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets," in *144th AES Convention, paper 9987, Milan, Italy*, May 2018.

[82] P. Coleman, A. Franck, P. J. B. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, "Object-based reverberation for spatial audio," *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 66–77, 2017.

[83] J. M. Jot and K. S. Lee, "Augmented reality headphone environment rendering," in *Int. AES Conf. on Audio for Virtual & Augmented Reality, Los Angeles, CA, USA*, 2016.

[84] C. Pörschmann and A. Zebisch, "Psychoacoustic investigations on synthetically created diffuse reverberation," in *27th Tonmeistertagung - VDT Int. Convention*, Cologne, Germany, 2012.

[85] J. M. Arend, S. V. A. Garí, C. Schissler, F. Klein, and P. W. Robinson, "Six-degrees-of-freedom parametric spatial audio based on one mono room impulse response," *J. Audio Eng. Soc. 69(7/8):557-75*, 2021. [Online]. Available: https://doi.org/10.17743/jaes.2021.0009

[86] M. Crocco and A. Del Bue, "Room impulse response estimation by iterative weighted l1-norm," in *23rd European Signal Processing Conference (EUSIPCO), 2015*, 2015, pp. 1895–1899.

[87] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[88] F. Xiong, S. Goetze, B. Kollmeier, and B. T. Meyer, "Exploring auditory-inspired acoustic features for room acoustic parameter estimation from monaural speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1809–1820, 2018.

[89] S. Li, R. Schlieper, and J. Peissig, "A hybrid method for blind estimation of frequency dependent reverberation time using speech signals," in *IEEE ICASSP, Brighton, UK, 2019*, 2019, pp. 211–215.

[90] P. Zahorik, "Spatial Hearing in Rooms and Effects of Reverberation," in *Binaural Hearing*, R. Litovsky, M. Goupell, R. R. Fay, and A. Popper, Eds. Springer Handbook of Auditory Research, 2021, pp. 243–280.

[91] M. Vorländer, "What do we know in room acoustics?" in *Forum Acusticum, Aalborg, Denmark*, 2011.

[92] G. T. Fechner, *Elemente der Psychophysik*, 2nd ed. Breitkopf u. Härtel, 1889.

[93] S. Klockgether and S. van de Par, "Just noticeable differences of spatial cues in echoic and anechoic acoustical environments," *J. Acoust. Soc. Am., 140(4), pp. EL352-EL357*, 2016.

[94] F. Martellotta, "The just noticeable difference of center time and clarity index in large reverberant spaces," *J. Acoust. Soc. Am. 128(2), pp. 654–663*, 2010.

[95] F. d. S. Dorrego and M. Vigeant, "A study of the just noticeable differences of early decay time for symphonic halls," *J. Acoust. Soc. Am. 151(1) Jan., pp. 80-94*, 2022.

[96] S. Weinzierl and M. Vorländer, "Room acoustical parameters as predictors of room acoustical impression: What do we know and what would we like to know?" *Acoustics Australia, 43(1), pp. 41-48*, 2015.

[97] J. S. Bradley, "Review of objective room acoustics measures and future needs," *Applied Acoustics*, vol. 72, no. 10, pp. 713–720, Oct. 2011.

[98] J. van Dorp Schuitman, D. De Vries, and A. Lindau, "Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model," *J. Acoust. Soc. Am., 133(3), pp. 1572-1585*, 2013.

[99] F. E. Toole, *Sound reproduction - The Acoustic and Psychoacoustics of Loudspeaker and Rooms*. 3rd edition, New York, London: Routledge, 2017.

[100] N. Kaplanis, S. Bech, T. Lokki, T. van Waterschoot, and S. Holdt Jensen, "Perception and preference of reverberation in small listening rooms for multi-loudspeaker reproduction," *J. Acoust. Soc. Am., 146(5), pp. 3562-3576*, 2019.

[101] P. Zahorik, "Perceptually relevant parameters for virtual listening simulation of small room acoustics," *J. Acoust. Soc. Am., 126 (2), Aug.*, 2009.

[102] J. Buchholz, J. Mourjopoulos, and J. Blauert, "Room masking: Understanding and modelling the masking of reflections in rooms," in *110th AES Convention, paper 5312, Amsterdam, The Netherlands*, 2001.

[103] X. Zhong, W. Guo, and J. Wang, "Audible threshold of early reflections with different orientations and delays," *Sound & Vibration, Dec.*, 2018.

[104] D. R. Begault, B. U. McClain, and M. R. Anderson, "Early reflection thresholds for anechoic and reverberant stimuli within a 3-D Sound Display," in *18th International Congress on Acoustics (ICA), Kyoto, Japan*, 2004.

[105] J. Buchholz, "Characterizing the monaural and binaural process underlying reflection masking," *Hearing Research, Vol . 232, pp. 52-66*, 2007.

[106] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization," *The American Journal of Psychology*, vol. 62, pp. 315–336, 1949.

[107] H. Haas, "Über den Einfluß eines Einfachechos auf die Hörsamkeit von Sprache," *Acustica 1(2), pp. 49-58*, 1951.

[108] ——, "The influence of a single echo on the audibility of speech," *J. Audio Eng. Soc. 20(2), pp. 146-159*, 1972.

[109] A. D. Brown, G. C. Stecker, and D. J. Tollin, "The precedence effect in sound localization," *J. Assoc. Res. Otolaryngol.*, vol. 16, no. 1, pp. 1–28, 2015.

[110] P. M. Zurek, "The precedence effect," in *Directional Hearing*, W. Yost and G. Gourevitch, Eds. Springer Verlag, New York, 1987, pp. 85–105.

[111] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654, 1999.

[112] P. W. Robinson, A. Walther, C. Faller, and J. Braasch, "Echo thresholds for reflections from acoustically diffusive architectural surfaces," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 2755–2764, 2013.

[113] F. Wendt and H. R., "Precedence effect for specular and diffuse reflections," *Acta Acustica,5(1), 2021, doi: 10.1051/aacus/2020027.*, 2021.

[114] J. Paasonen, A. Karapetyan, J. Plogsties, and V. Pulkki, "Proximity of surfaces—acoustic and perceptual effects," *J. Audio Eng. Soc.*, vol. 65, no. 12, pp. 997–1004, Dec. 2017.

[Online]. Available: http://dx.doi.org/10.17743/jaes.2017.0039

[115] M. Barron, "Subjective effects of first reflections in concert halls - the need for lateral reflections," *Journal of Sound and Vibration 15(4):475-494*, 1971.

[116] S. Brunner, H.-J. Maempel, and S. Weinzierl, "On the audibility of comb filter distortions," in *112nd AES Convention, paper 7047, Vienna, Austria*, 2007.

[117] F. Bilsen and R. Ritsma, "Repetition pitch and its implication for hearing theory," *Acta Acustica united with Acustica, Vol. 22 (2), pp. 63-73, Jan.*, 1969.

[118] K. Roman, H.-J. Maempel, and S. Weinzierl, "Zur Wahrnehmung überlagerter Signalreflexionen [About the perception of superposed signal reflections]," in *34th Annual Conference on Acoustics, Dresden*, 2008.

[119] S. Bech, "Timbral aspects of reproduced sound in small rooms. i," *J. Acoust. Soc. Am.*, vol. 97, no. 3, pp. 1717–1726, Mar. 1995.

[120] ——, "Timbral aspects of reproduced sound in small rooms. ii," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3539–3549, 1996.

[121] F. Zotter and M. Frank, "Investigation of auditory objects caused by directional sound sources in rooms," *Acta Physica Polonica A*, vol. 128, no. 1, pp. A5–A10, 2015.

[122] H. Steffens, S. van de Par, and S. D. Ewert, "The role of early and late reflections on perception of source orientation," *J. Acoust. Soc. Am. 149(4). April 2021. doi: 10.1121/10.0003823*, 2021.

[123] S. Bech, "Spatial aspects of reproduced sound in small rooms," *J. Acoust. Soc. Am. 103(1):434-45*, 1998.

[124] M. Barron and A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," *Journal on Sound and Vibration*, vol. 77, no. 2, pp. 211–232, 1981.

[125] T. Okano, L. L. Beranek, and T. Hidaka, "Relations among interaural cross-correlation coefficient ($IACC_E$), lateral fraction (LF_E), and apparent source width (ASW) in concert halls," *J. Acoust. Soc. Am.*, vol. 104, no. 1, pp. 255–265, 1998.

[126] D. Johnson and H. Lee, "Perceptual threshold of apparent source width in relation to the azimuth of a single reflection," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. EL272–EL276, 2019.

[127] B. Shinn-Cunningham and S. Ram, "Identifying where you are in the room: Sensitivity to room acoustics," in *Int Conf. on Auditory Display, Boston, MA, USA*, 2003.

[128] A. J. Kolarik, A. C. Scarfe, B. C. Moore, and S. Pardhan, "Blindness enhances auditory obstacle circumvention: Assessing echolocation, sensory substitution, and visual-based navigation," *PloS ONE*, vol. 12, no. 4, p. e0175750, 2017.

[129] F. Klein, A. Neidhardt, M. Seipel, and T. Sporer, "Training on the acoustical identification of the listening position in a virtual environment," in *143th AES Convention, paper 9834, New York, USA*, 2017.

[130] C. Schneiderwind and A. Neidhardt, "Perceptual differences of position dependent room acoustics in a small conference room," in *Int. Symposium on Room Acoustics (ISRA), Amsterdam, Netherlands*, 2019.

[131] T. Welti and R. E. Jensen, "The importance of reflections in a binaural room impulse response," in *114th AES Convention, Amsterdam, The Netherlands, 2003*, 2003.

[132] H. Hacıhabiboğlu and F. Murtagh, "Perceptual simplification for model-based binaural room auralisation," *Applied Acoustics*, vol. 69, no. 8, pp. 715–727, 2008.

[133] R. R. Torres, P. U. Svensson, and M. Kleiner, "Computation of edge diffraction for more accurate room acoustic auralization," *J. Acoust. Soc. Am., 109(2), 600–610*, 2001. [Online].

Available: https://doi.org/10.1121/1.1340647

[134] P. Calamia, "Advances in edge diffraction," Ph.D. dissertation, Princeton University, 2009.

[135] B. G. Shinn-Cunningham, "Learning reverberation: Considerations for spatial auditory displays," in *Int. Conf. on Auditory Display (ICAD), Atlanta, Georgia, USA, April*, 2000.

[136] F. Wendt, F. Zotter, M. Frank, and R. Höldrich, "Auditory distance control using a variable-directivity loudspeaker," *Applied Sciences, 7(7), p. 666*, 2017.

[137] M.-V. Laitinen, A. Politis, I. Huhtakallio, and V. Pulkki, "Controlling the perceived distance of an auditory object by manipulation of loudspeaker directivity," *J. Acoust. Soc. Am.*, vol. 137, no. 6, pp. EL462–EL468, 2015.

[138] D. Cabrera, "Acoustic clarity and auditory room size perception," in *14th Int. Congress on Sound & Vibration, Cairns, Australia, 2007*, 2007, pp. 9–12.

[139] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," *J. Acoust. Soc. Am. 124 (1), pp. 450-461, July*, 2008. [Online]. Available: https://doi.org/10.1121/1.2936368

[140] P. Zahorik, E. Brandewie, and V. P. Sivonen, "Auditory perception in reverberant sound fields and effects of prior listening exposure," in *Principles and Applications of Spatial Hearing*. World Scientific, 2011, ch. 2, pp. 24–34.

[141] R. K. Clifton, R. L. Freyman, R. Y. Litovsky, and D. McCall, "Listeners' expectations about echoes can raise or lower echo threshold," *J. Acoust. Soc. Am., 95(3), pp. 1525-1533*, 1994.

[142] R. Keen and R. L. Freyman, "Release and re-buildup of listeners' models of auditory space," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3243–3252, 2009.

[143] S. Clapp and B. Seeber, "Auditory room learning and adaptation to sound reflections," in *The Technology of Binaural Understanding*. Springer Int. Publishing, 2020.

[144] B. U. Seeber, M. Müller, and F. Menzer, "Does learning a room's reflections aid spatial hearing?" in *22nd Int. Congress on Acoustics, Buenos Aires, Argentina*, Sep. 2016.

[145] K. Brandenburg, F. Klein, A. Neidhardt, U. Sloma, and S. Werner, "Creating auditory illusions with binaural technology," in *The Technology of Binaural Understanding*. Springer Int. Publishing, 2020, pp. 623–663. [Online]. Available: https://doi.org/10.1007/978-3-030-00386-9_21

[146] F. Klein, S. Werner, and T. Mayenfels, "Influences of training on externalization of binaural synthesis in situations of room divergence," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 178–187, 2017.

[147] P. Zahorik, "Adaptation to room acoustics and its effect on speech understanding," in *23rd Int. Congress on Acoustics, Aachen, Germany*, 2019.

[148] F. Klein, S. Werner, G. Götz, and K. Brandenburg, "Auditory adaptation in real and virtual rooms," in *Int. Symp. on Auditory and Audiological Research*, Nyborg, Denmark, 2019, 2019, pp. 341–348.

[149] D. Romblom, C. Guastavino, and P. Depalle, "Perceptual thresholds for non-ideal diffuse field reverberation," *J. Acoust. Soc. Am.*, vol. 140, no. 5, pp. 3908–3916, 2016.

[150] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of model and signal-based predictors of the mixing time in binaural room impulse responses," *J. Audio Eng. Soc., 60(11):887-898*, 2012.

[151] K. Meesawat and D. Hammershøi, "The time when the reverberation tail in a binaural room impulse response begins," in *115th AES Convention, paper 5859, New York, NY, USA*, 2003.

[152] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *PNAS Plus, Psychological and Cognitive Sciences.*, 2016.

[153] S. Djordjević, H. Hacihabiboğlu, Z. Cvetković, and E. De Sena, "Evaluation of the perceived naturalness of artificial reverberation algorithms," in *148th AES Convention, Online, 2020*, Jun. 2020.

[154] E. De Sena, H. Hacihabiboğlu, and Z. Cvetković, "Scattering delay network: An interactive reverberator for computer games," in *41st Int. Conf. Audio for Games. London, UK*, 2011.

[155] A. Vecchi, A. Kohlrausch, W. Lachenmayr, and E. Mommertz, "Predicting the perceeived reverberation in different room acoustic environments using a binaural auditory model," *J. Acoust. Soc. Am. 141 (4), April*, 2017.

[156] G. A. Soulodre, "New objective measures of listener envelopment," in *18th Int. Congress on Acoustics*, Kyoto, Japan, 2004, 2004, pp. 2465–2468.

[157] L. L. Beranek, "Listener envelopment LEV, strength G and reverberation time RT in concert halls," in *20th Int. Congress on Acoustics*, Sydney, Australia, 2010, 2010.

[158] B. Alary, P. Massé, S. J. Schlecht, M. Noisternig, and V. Välimäki, "Perceptual analysis of directional late reverberation," *J. Acoust. Soc. Am.*, vol. 149, no. 5, pp. 3189–3199, 2021.

[159] V. O. Knudsen, "Resonance in small rooms," *J. Acoust. Soc. Am.*, vol. 4, no. 1A, pp. 20–37, 1932.

[160] M. R. Schroeder, "The "Schroeder frequency" revisited," *J. Acoust. Soc. Am.*, vol. 99, no. 5, pp. 3240–3241, 1996.

[161] M. Skålevik, "Schroeder frequency revisited," in *Proceedings of Forum Acusticum, Aalborg, Denmark, 2011*, 2011.

[162] M. Karjalainen, P. Antsalo, A. Mäkivirta, and V. Välimäki, "Perception of temporal decay of low-frequency room modes," in *116th AES Convention, Berlin, Germany, 2004*, May 2004.

[163] S. Bech, "Perception of timbre of reproduced sound in small rooms: Influence of room and loudspeaker position," *J. Audio Eng. Soc.*, vol. 42, no. 12, pp. 999–1007, Dec. 1994.

[164] S. E. Olive, P. L. Schuck, S. L. Sally, and M. E. Bonneville, "The effects of loudspeaker placement on listener preference ratings," *J. Audio Eng. Soc.*, vol. 42, no. 9, pp. 651–669, Sep. 1994.

[165] R. Bücklein, "The audibility of frequency response irregularities," *J. Audio Eng. Soc.*, vol. 29, no. 3, pp. 126–131, Mar. 1981.

[166] B. M. Fazenda, M. Stephenson, and A. Goldberg, "Perceptual thresholds for the effects of room modes as a function of modal decay," *J. Acoust. Soc. Am.*, vol. 137, no. 3, pp. 1088–1098, Mar. 2015.

[167] B. Moore, *An Introduction to the Psychology of Hearing*. BRILL, 6th edition, 2013.

[168] K. McAnally and R. Martin, "Sound localization with head movement: Implications for 3-d audio displays," *Frontiers in Neuroscience*, 2014.

[169] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization," *Journal of Experimental Psychology, Vol. 27, No. 4, Oct.*, 1940.

[170] W. Simpson and L. Stanton, "Head movement does not facilitate perception of the distance of a source of sound," *American Journal of Psychology, 86, pp. 151-159*, 1973.

[171] R. A. Epstein, E. Z. Patai, J. B. Julian, and H. J. Spiers, "The cognitive map in humans: Spatial navigation and beyond," *Nature Neuroscience*, vol. 20, no. 11, pp. 1504–1513, 2017.

[172] S. M. Weisberg and N. S. Newcombe, "Cognitive maps: Some people make them, some people struggle," *Current Directions in Psychological Science, 27(4), pp. 220-226*, 2018. [Online]. Available: https://doi.org/10.1177/0963721417744521

[173] F. Durgin *et al.*, "Self-motion perception during locomotor recalibration: More than meets the eye," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 31, pp.

398–419, 2005.

[174] F. H. Durgin, "When walking makes perception better," *Current Directions in Psychological Science, 18(1), pp. 43-47*, 2009. [Online]. Available: https://doi.org/10.1111/j.1467-8721.2009.01603.x

[175] C. S. Harris, "Insight or out of sight? two examples of perceptual plasticity in the human adult," in *Visual coding and adaptability*. Hillsdale, NJ: Erlbaum, 1980, pp. 95–149.

[176] A. Väljamäe, "Auditorily-induced illusory self-motion: A review," *Brain Research Reviews, Vol. 61, No. 2, pp. 240-255*, 2009.

[177] A. Väljamäe, P. Larsson, D. Västfjäll, and M. Kleiner, "Travelling without moving: Auditory scene cues for translational self-motion," in *International Conference on Auditory Display, Limerick, Ireland*, 2005.

[178] S. Carlile and J. Leung, "The Perception of Auditory Motion," *Trends in Hearing, Vol. 20, pp. 1-19*, 2016. [Online]. Available: https://doi.org/10.1177/2331216516644254

[179] D. Genzel, M. Schutte, W. Brimijoin, P. R. MacNeilage, and L. Wiegrebe, "Psychophysical evidence for auditory motion parallax," *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 2018.

[180] M. Wexler and J. J. van Boxtel, "Depth perception by the active observer," *Trends in Cognitive Sciences, 9(9), pp. 431–438*, 2005. [Online]. Available: https://doi.org/10.1016/j.tics.2005.06.018

[181] B. Rogers and M. Graham, "Motion parallax as an independent cue for depth perception," *Perception*, vol. 8, pp. 125–134, 1979.

[182] K. Hayashibe, "The efficient range of velocities for inducing depth perception by motion parallax," *Percept Mot Skills. Oct; 83(2):659-74*, 1996.

[183] D. Ashmead, D. Davis, and A. Northington, "Contribution of Listeners' Approaching Motion to Auditory Perception," *Journal of Experimental Psychology, Human Perception and Performance, Vol. 21, No. 2, 239-256*, 1995.

[184] J. Speigle and J. Loomis, "Auditory Distance Perception by Translating Observers," *IEEE Research Properties in Virtual Reality Symposium, San Jose, CA, USA*, 1993.

[185] A. J. Kolarik, S. Cirstea, S. Pardhan, and B. C. Moore, "A summary of research investigating echolocation abilities of blind and sighted humans," *Hearing Research*, vol. 310, pp. 60–68, 2014.

[186] L. D. Rosenblum, M. S. Gordon, and L. Jarquin, "Echolocating distance by moving and stationary listeners," *Ecological Psychology, 12 (3), 181-206*, 2000.

[187] C. Störig and C. Pörschmann, "Investigations into velocity and distance perception based on different types of moving sound sources with respect to auditory virtual environments," *Journal of Virtual Reality and Broadcasting*, vol. 10, no. 4, 2013.

[188] A. Neidhardt, "Distance perception in virtual auditory environments with a moving avatar," in *42nd Annual Conference on Acoustics, Aachen, Germany*, 2016.

[189] M. Boerum, B. Martin, R. King, and G. Massenburg, "Lateral listener movement on the horizontal plane: Part 2 sensing motion through binaural simulation in a reverberant environment," in *Int. AES Conf. on Audio for Virtual & Augm. Reality, Los Angeles, CA, USA*, 2016.

[190] G. Kearney, X. Liu, A. Manns, and M. Gorzel, "Auditory Distance Perception with Static and Dynamic Binaural Rendering," in *AES 57th Int. Conference, Hollywood, CA, USA*, 2015.

[191] S. Werner, F. Klein, A. Neidhardt, U. Sloma, C. Schneiderwind, and K. Brandenburg, "Creation of Auditory Augmented Reality using a position-dynamic binaural synthesis system—Technical components, psychoacoustic needs, and perceptual evaluation," *Applied Sciences 11(3),*

Jan., p. 1150, 2021. [Online]. Available: https://doi.org/10.3390/app11031150

[192] J. G. Tylka and E. Y. Choueiri, "Fundamentals of a parametric method for virtual navigation within an array of ambisonics," *J. Audio Eng. Soc. 68(3),*, pp. 120–137, Mar. 2020.

[193] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. P. Habets, "Six-degrees-of-freedom binaural reproduction of first order ambisonics with distance information," in *AES Int. Conference on Audio for Virtual and Augmented Reality, Redmond, WA, USA*, 08 2018.

[194] S. E. Olive and F. E. Toole, "The detection of reflection in typical rooms," *J Audio Eng Soc 37(7/8), Jul/Aug 1989*, 1989.

[195] P. Stade, "Perzeptive Untersuchung zur Mixing Time und deren Einfluss auf die Auralisation [Perceptual investigation on the mixing time and its influence on the auralization]," in *41st Annual Conference on Acoustics, Nuremberg, Germany*, 2015.

[196] B. Rakerd, W. M. Hartmann, and J. Hsu, "Echo suppression in the horizontal and median sagittal planes," *J Acoust Soc Am 107(2):1061-1064, 2000.*, 2000.

[197] M. A. Gerzon, "Ambisonics in Multichannel Broadcasting and Video," *J. Audio. Eng. Soc.*, vol. 33, no. 11, pp. 859–871, 1985.

[198] F. Zotter and M. Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*, 1st ed. Springer Int. Cham, Vol. 19 of Springer Topics in Signal Processing, 2019, ISBN-13: 978-3030172060.

[199] I. Engel, H. Craig, S. Garí, P. Robinson, and L. Picinali, "Perceptual implications of different ambisonics-based methods for binaural reverberation," *J. Acoust. Soc. Am.*, vol. 149, no. 2, pp. 895 ff, 2021.

[200] E. Patricio, A. Rumiński, A. Kuklansiński, Ł. Januszkiewicz, and T. Żernicki, "Toward six degrees of freedom audio recording and playback using multiple ambisonics sound fields," in *146th Int. AES Convention, Dublin, Ireland, paper 10141*, 2019.

[201] F. Heller, J. Jevanesan, P. Dietrich, and J. Borchers, "Where are we? Evaluating the current rendering fidelity of mobile audio augmented reality systems," in *18th Int. Conf. MobileHCI, Florence, Italy*, 2016.

[202] M. Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," *Phil. Trans. R. Soc. B., 364, 3549–3557*, 2009.

[203] ——, "Immersion and the illusion of presence in virtual reality," *British Journal of Psychology*, 2018.

[204] C. Kuhn-Rahloff, "Prozesse der Plausibilitätsbeurteilung am Beispiel ausgewählter elektroakustischer Wiedergabesituationen," Ph.D. dissertation, TU Berlin, Germany, 2011.

[205] M. E. Latoschik and C. Wienrich, "Congruence and Plausibility, not Presence: Pivotal conditions for XR experiences and effects, a novel approach," *Front. Virtual Real., Sec. Virtual Reality and Human Behaviour, 16 June*, 2021. [Online]. Available: https://doi.org/10.3389/frvir.2022.694433

[206] R. Skarbez, F. B. Jr., and M. Whitton, "A survey of presence and related concepts," *ACM Computing Surveys 50(6), 96:1-39*, 2017.

[207] A. Moore, A. Tew, and R. Nicol, "An initial validation of individualized crosstalk cancellation filters for binaural perceptual experiments," *J. Audio Eng. Soc.*, vol. 58, pp. 36–45, 2010. [Online]. Available: http://www.frontiersin.org/Journal/10.3389/fnins.2013.12345/abstract

[208] B. Maseiro, "Individualized binaural technology. measurement, equalization and perceptual evaluation," Ph.D. dissertation, RWTH Aachen, Germany, 2012.

[209] J. Oberem, B. Maseiro, and J. Fels, "Experiments on authenticity and plausibility of binaural reproduction ia headphones employing different recording methods," *Applied Acoustics*, vol.

114, pp. 71–78, 2016. [Online]. Available: http://dx.doi.org/10.1012/j.apacoust.2016.07.009

[210] V. Erbes, F. Schultz, A. Lindau, and S. Weinzierl, "An extraaural headphone system for optimized binaural reproduction," in *Fortschritte der Akustik: Tagungsband der 38. DAGA, Darmstadt, pp. 313-314*, 2012.

[211] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments," *Acta Acustica united with Acustica, Vol. 98, No. 5, pp. 804–810*, 2012. [Online]. Available: dx.doi.org/10.3813/AAA.918562

[212] W. M. Hartmann and A. Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am. 99(6), pp. 3678-3688*, 1996.

[213] E. Langendijk and A. Bronkhorst, "Fidelity of three-dimensional sound reproduction using a virtual auditory display," *J. Acoust. Soc. Am. 107(1), 528-537*, 2000.

[214] A. Lindau, T. Hohn, and S. Weinzierl, "Binaural resynthesis for comparative studies of acoustical environments," in *122nd AES Convention, paper 7032, Vienna, Austria*, 2007.

[215] K. Enge, M. Frank, and R. Höldrich, "Listening experiment on the plausibility of acoustic modeling in virtual reality," in *46th Annual Meeting on Acoustics (DAGA), Hannover, Germany*, 2020.

[216] M. Hofer, T. Hartmann, A. Eden, R. Ratan, and L. Hahn, "The role of plausibility in the experience of spatial presence in virtual environments," *Front. VR., April*, 2020.

[217] S. A. Wirler, N. Meyer-Kahlen, and S. J. Schlecht, "Towards Transfer-Plausibility for evaluating Mixed Reality audio in complex scenes," in *AES Int. Conference on Audio for Virtual and Augmented Reality, Online*, 2020.

[218] G. C. Stecker, T. M. Moore, M. Folkerts, D. Zotkin, and R. Duraiswami, "Towards objective measures of auditory co-immersion in virtual and augmented reality," in *AES Int. Conf. Audio for Virtual and Augm. Reality, Redmond, WA, USA*, 2018.

[219] M. Geier, J. Ahrens, and S. Spors, "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," in *124th AES Convention, paper 7330, Amsterdam, The Netherlands*, 2008.

[220] M. Geier and S. Spors, "Spatial Audio with the SoundScape Renderer," in *27th Tonmeistertagung - VDT Int. Convention, Cologne, Deutschland*, 2012.

[221] A. Vazquez Giner, "Scale - Conducting psychoacoustic experiments with dynamic binaural synthesis," in *Proceedings of the 41st DAGA, pp. 1128-1130*, 2015.

[222] A. Baskind, J.-C. Messonnier, and J.-M. Lyzwa, "Bipan: an experimental mixing tool for 3d-audio over headphones with open-source head tracker," in *29th Tonmeistertagung - VDT International Convention, Cologne, Germany*, 2016.

[223] F. Alouges and et al., "Mybino - binaural monitoring vst-plugin," last accessed on 27th March, 2017, http://www.cmap.polytechnique.fr/xaudio/mybino/.

[224] Firelight Technologies, "Fmod," last accessed on 27th March, 2017, http://www.fmod.com/.

[225] M. Kronlachner, "ambix ambisonics suite," 2014, last accessed on 27th March, 2017. [Online]. Available: http://www.matthiaskronlachner.com/?p=2015

[226] H. Gomersall, "pyFFTW Python module," last accessed on 27th March, 2017, https://github.com/pyFFTW/pyFFTW/.

[227] H. Pham, "pyAudio Python module," last accessed on 27th March, 2017, https://people.csail.mit.edu/hubert/pyaudio/.

[228] P. Bartz and et al., "Razor ahrs tracking modul," last accessed on 27th March, 2017, https://github.com/Razor-AHRS/razor-9dof-ahrs.

[229] P. Majdak, F. Zotter, F. Brinkmann, J. De Muynke, M. Mihocic, and M. Noisternig, "Spatially

oriented format for acoustics 2.1: Introduction and recent advances," *J. Audio Eng. Soc., 70(7/8), pp. 565-584*, 2022.

[230] GENELEC, "Genelec 1030a monitoring system - data sheet," Genelec, Tech. Rep., 2003, data Sheet.

[231] ——, "Photo 1030a monitoring system," *Unknown Journal*, 2018, last visit 19th Nov. [Online]. Available: https://www.genelec.com/support-technology/previous-models/1030a-studio-monitor

[232] I. The International Telecommunication Union-Radiocommunication Assembly, "ITU-R BS.1116-3: Methods for the subjective assessment of small impairments in audio systems," Place des Nations, Genève, Suisse, Tech. Rep., 2015.

[233] A. Wabnitz, N. Epain, C. Jin, and A. van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Int. Symp. on Room Acoustics, ISRA, Melbourne, Australia*, 2010. [Online]. Available: https://github.com/Andronicus1000/MCRoomSim/blob/master/MCRoomSim.m

[234] B. Bernschütz, "A spherical far field HRIR/HRTF compilation," in *39. Jahrestagung für Akustik, Meran*, 2013.

[235] PBTech, "Photo HTC Vive," *Unknown Journal*, 2018, last 19th Nov. [Online]. Available: https://www.pbtech.co.nz/product/MVRHTC1001/HTC-VIVE-Consumer-Edition-Virtual-Reality-Headset

[236] C. Kuhn-Rahloff, "Realitätstreue, Natürlichkeit, Plausibilität: Perzeptive Beurteilungen in der Elektroakustik," Ph.D. dissertation, TU Berlin, Germany, 2012.

[237] N. Knoop, "Orientierung in einem virtuellen Raum mit beweglichem Avatar [Orientation in a virtual room with a dynamic avatar]," Master's thesis, Technische Universität Ilmenau, 2016.

[238] F. Wefers, J. Stienen, S. Pelzer, and M. Vorländer, "Interactive acoustic virtual environment using distributed room acoustic simulations," in *EAA Joint Symp. on Auralization and Ambisonics, Berlin, Germany*, 2014.

[239] S. W. Clapp and B. U. Seeber, "Sound localization in partially updated room auralizations," in *42nd Annual Meeting on Acoustics, Aachen, Germany*, 2016.

[240] A. Franck, "Efficient frequency-domain filter crossfading for fast convolution with application to binaural synthesis," in *AES 55th International Conference, Helsinki, Finland*, 2014.

[241] C. Mittag, S. Werner, and F. Klein, "Development and evaluation of methods for the synthesis of binaural room impulse responses based in spatially sparse measurements in real rooms," in *43rd Annual Conference on Acoustics, Kiel, Germany*, 2017.

[242] C. Pörschmann and P. Stade, "Auralizing listener position shifts of measured room impulse responses," in *42nd Annual Conference on Acoustics, Aachen, Germany*, 2016.

[243] S. R. Narum, "Beyond Bonferroni: Less conservative analysis for conservation genetics," *Conservation Genetics*, pp. 7:783–787, 2006.

[244] D. Begault and E. Wenzel, "Direct Comparison of the Impact of Head Tracking, Reverberation and Individualized Head-related Transfer Functions on the Spatial Perception of the Virtual Speech Source," *J. Audio Eng. Soc., Vol. 49, No. 10, October*, 2001.

[245] ITU, "ITU-R BS.1284-2: General methods for the subjective assessment of sound quality," (The International Telecommunication Union-Radiocommunication Assembly), Place des Nations, Genève, Suisse, Tech. Rep., 2019.

[246] J. Udesen, T. Piechowiak, and F. Gran, "The effect of vision on psychoacoustic testing with headphone-based virtual sound," *J. Audio Eng. Soc., Vol. 63, No. 7/8, pp. 552–561*, Jul. 2015. [Online]. Available: 10.17743/jaes.2015.0061

[247] S. Werner, G. Götz, and F. Klein, "Influence of head tracking on the externalization of auditory

events at divergence between synthesized and listening room using a binaural headphone system," in *142nd AES Convention, paper 9690, Berlin, Germany*, May 2017.

[248] D. Perrott and K. Saberi, "Minimum audible angle thresholds for sources varying in both elevation and azimuth," *J. Acoust. Soc. Am. 87 (4)*, pp. 1728–1731, 1990. [Online]. Available: https://doi.org/10.1121/1.399421

[249] M. Lundbeck, G. Grimm, V. Hohmann, S. Laugesen, and T. Neher, "Sensitivity to Angular and Radial Source Movements as a Function of Acoustic Complexity in Normal and Impaired Hearing," *Trends in Hearing, Vol. 21, pp. 1-14*, 2017. [Online]. Available: https://doi.org/10.1177/2331216517717152

[250] K. Saberi and D. R. Perrott, "Minimum audible movement angles as a function of sound source trajectory," *J. Acoust. Soc. Am. 88 (6), Dec.*, 1990. [Online]. Available: https://doi.org/10.1121/1.399984

[251] W. O. Brimijoin and M. A. Akeroyd, "The moving minimum audible angle is smaller during self-motion than during source motion," *Frontiers in Neuroscience, Vol. 8, 273*, 2014. [Online]. Available: https://doi.org/10.3389/fnins.2014.00273

[252] O. Rummukainen, S. Schlecht, and E. Habets, "Self-translation induced minimum audible angle," *J. Acoust. Soc. Am. 144(4), EL340, Oct.*, 2018. [Online]. Available: https://doi.org/10.1121/1.5064957

[253] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J. Acoust. Soc. Am., Vol. 112, No. 5, p. 2110-2117*, 2002. [Online]. Available: https://doi.org/10.1121/1.1506692

[254] M. Florentine and S. Buus, "An excitation-pattern model for intensity discrimination," *J. Acoust. Soc. Am. ,70(6), pp. 1646-1654*, 1981. [Online]. Available: https://doi.org/10.1121/1.387219

[255] Z. L. Chen, G. S. Hue, B. R. Glasberg, and B. C. J. Moore, "A new method of calculating auditory excitation patterns and loudness for steady sounds," *Hearing Research, 282, 204-215*, 2011. [Online]. Available: https://doi.org/10.1016/j.heares.2011.08.001

[256] A. Lindau, H.-J. Maempel, and S. Weinzierl, "Minimum BRIR grid resolution for dynamic binaural synthesis," in *Acoustics '08, Paris, pp. 3851-3856*, 2008.

[257] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology, 140, 52*, 1932.

[258] J. de Winter and D. Dodou, "Five-point likert items: t-test versus mann-whitney-wilcoxon," *Practical Assessment, Research & Evaluation, Vol. 15, 11, October*, 2010.

[259] T. Ajdler, L. Sbiaz, and M. Vetterli, "The Plenacoustic Function and Its Sampling," *IEEE Transactions on Signal Processing, Vol 54, No. 10, October*, 2006.

[260] S. Werner, F. Klein, and G. Götz, "Investigation on spatial auditory perception using non-uniform spatial distribution of binaural room impulse responses," in *5th Int. Conf. on Spatial Audio (ICSA), Ilmenau, Germany*, 2019.

[261] D. Satongar, C. Pike, Y. W. Lam, A. I. Tew *et al.*, "The influence of headphones on the localization of external loudspeaker sources," *J. Audio Eng. Soc., Vol. 63, No. 10, pp. 799–810*, Oct. 2015.

[262] C. Pike, F. Melchior, and T. Tew, "Assessing the plausibility of non-individualised dynamic binaural synthesis in a small room," in *55th Int. AES Conference on Spatial Audio, Helsinki, Finland*, 2014.

[263] M. Herzog, G. Francis, and A. Clarke, *Understanding Statistics and Experimental Design - How to not lie with Statistics*. Germany: Springer Open Books - Learning Materials in Biosciences, 2019.

[264] N. Macmillan and C. Creelman, *Detection theory - a user's guide*, 2nd ed. Psychology Press,

2004.

[265] T. Wickens, *Elementary Signal Detection Theory*. Germany: Oxford University Press, USA, 2001.

[266] M. Hautus, "Corrections for extreme proportions and their biasing effects on estimated values of d'," *Behavior Research Methods, Instruments, & Computers 27, 46–51*, 1995.

[267] C. J. Clopper and R. S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika Vol. 26, No. 4 (Dec., 1934), pp. 404-413*, 1934.

[268] C. Pörschmann, J. Arend, and R. Gillioz, "How wearing headgear affects measured head-related transfer functions," in *1st EAA Spatial Audio Signal Processing Symposium, Paris, France*, 2019.

[269] G. Plenge, "On the problem of "in head localization"," *Acta Acustica united with Acustica, 26(5), 241-252(12)*, 1972.

[270] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests - a review," *J. Audio Eng. Soc., Vol. 56, No. 6, pp. 427–451*, 2008.

[271] ITU, "ITU-T P.800.2," ITU, Tech. Rep., 2016.

[272] J. S. Abel and P. Huang, "A simple, robust measure of reverberation echo density," in *121st Int. AES Convention, San Francisco, CA, USA*, 10 2006.

[273] F. Brinkmann and S. Weinzierl, "AKtools – an open software toolbox for signal acquisition, processing, and inspection in acoustics," in *142nd AES Convention, Berlin, Germany, e-Brief 309*, 2017. [Online]. Available: https://www2.ak.tu-berlin.de/~akgroup/AKtools.zip

[274] S. Göring, R. Rao Ramachandra Rao, S. Fremerey, and A. Raake, "AVRate Voyager: an open source online testing platform," in *IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021. [Online]. Available: https://github.com/Telecommunication-Telemedia-Assessment/AVrateVoyager

[275] A. Neidhardt and A. M. Zerlik, "The availability of a real hidden reference affects the plausibility of position-dynamic auditory AR," *Frontiers in Virtual Reality*, 2021. [Online]. Available: https://doi.org/10.3389/frvir.2021.678875

[276] M. Gospodarek, O. Warusfel, P. Ripollés, and A. Roginska, "Methodology for perceptual evaluation of plausibility with self-translation of the listener," in *AES Conf. on Audio for Virtual and Augm. Reality, Redmond, WA, USA*, 2022.

[277] H. Schepker, F. Denk, B. Kollmeier, and S. Doclo, "Acoustic transparency in hearables—perceptual sound quality evaluations," *J. Audio Eng. Soc.*, vol. 68, no. 7/8, pp. 495–507, 2020.

[278] C. Schneiderwind, A. Neidhardt, and D. Meyer, "Comparing the effect of different open headphone models on the perception of a real sound source," in *150th AES Convention, paper 10489, Online*, May 2021.

[279] R. Gupta, R. Ranjan, J. He, W.-S. Gan, and S. Peksi, "Acoustic transparency in hearables for augmented reality audio: Hear-through techniques review and challenges," in *AES Int. Conf. on Audio for Virtual & Augm. Reality, San Francisco, CA, USA, 2020*, 2020.

[280] O. Doll, "Creation and evaluation of an AAR scenario using an electro-acoustically transparent hearpiece," Master's thesis, Technische Universität Ilmenau, 2022.

[281] F. Denk, M. Hiipakka, B. Kollmeier, and S. M. Ernst, "An individualized acoustically transparent earpiece for hearing devices," *International Journal of Audiology*, vol. 57, no. sup3, pp. S62–S70, 2018.

[282] N. Meyer-Kahlen and S. J. Schlecht, "Assessing room acoustic self-localization using a virtual blindfold," in *DAGA Wien*, 2021.

[283] N. Meyer-Kahlen, S. J. Schlecht, and T. Lokki, "Clearly audible room acoustical differences may not reveal where you are in a room," *J. Acoust. Soc. Am. 152 (2)*, pp. 877–887, Aug. 2022. [Online]. Available: https://doi.org/10.11

[284] M. von Berg, J. Steffens, S. Weinzierl, and D. Müllensiefen, "Assessing room acoustic listening expertise," *J. Acoust. Soc. Am. 150 (4), Oct*, 2021. [Online]. Available: https://doi.org/10.1121/10.0006574.

[285] T. Okano, "Image shift caused by strong lateral reflections, and its relation to inter-aural cross correlation," *J. Acoust. Soc. Am. 108(5), Pt. 1, Nov.*, 2000.

[286] M. T. Pastore and J. Braasch, "The precedence effect with increased lag level," *J. Acoust. Soc. Am. 138 (4), pp. 2079–2089*, 2015. [Online]. Available: https://doi.org/10.1121/1.4929940

[287] A. Neidhardt, "Data set and physical analysis: BRIRs and SRIRs for walking toward, past and behind virtual loudspeakers in two rooms," in *154th AES Convention, Espoo/Helsinki, Finland*, 2023.

[288] C. W. Pike, "Evaluating the perceived qualityof binaural technology," Ph.D. dissertation, University of York, Electronic Engineering, 2019.

[289] M. Mijić and D. Masovic, "Reverberation radius in real rooms," *Telfor Journal, Vol 2, No. 2 2010.*, 2010.

[290] M. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am. 37, 409*, 1965.

[291] D. Cabrera, J. Xun, and M. Guski, "Calculation of the reverberation time from the impulse response: A comparison of software implementations," *Acoust. Aust. 44:369-378, May*, 2016.

[292] M. Barron, "Using the standard objective measures for concert auditoria, iso 3382, to give reliable results," *Acoustical Science and Technology, (26)2:162-196*, 2005.

[293] F. D. S. Dorrego and M. C. Vigeant, "A study of the just noticeable difference of early decay time (EDT)," in *Auditorium Acoustics, Hamburg, Germany*, 2018, pp. 273–282.

[294] H. Kuttruff, *Room Acoustics - 5th edition*. Spon Press, 2009.

[295] H. Arau-Puchades and U. Berardi, "A revised sound energy theory based on a new formula for the reverberation radius in rooms with non-diffuse sound field," *Archives of Acoustics, Vol. 40, No. 1, pp. 33-40.*, 2015.

[296] G. von Békésy, "Über die Enstehung der Entfernungsempfindung beim Hören [On the origin of distance perception in hearing]," *Akust. Z. 3, pp. 31-41*, 1938.

[297] J.-C. Messonnier and A. Moraud, "Auditory distance perception: criteria and listening room," in *130th Int. Conv. of the AES, paper 8373, London, UK*, 2011.

[298] W. Reichardt and W. Schmidt, "Die hörbaren Stufen des Raumeindrucks bei Musik [The audible levels of spatial impression with music]," *Acustica*, vol. 17, pp. 175–179, 1966.

[299] E. Meyer and R. Thiele, "Room acoustical investigations in numerous concert halls and broadcasting studois using more recent measurement techniques," *Acustica 6:425.*, 1956.

[300] W. Reichardt, A. O. Abdel, and W. Schmidt, "Dependence of the limit between useful and useles transparency on the kind of music, of the reverberation time and the onset time of sound decay," *Applied Acoustics 7:243*, 1974.

[301] J. S. Bradley, "Predictors of speech intelligibility in rooms," *J. Acoust. Soc. Am. 80:837*, 1986.

[302] R. Kürer, "Isolation of single number criteria from impulse measurements in room acoustics," *Acustica 21:370*, 1969.

[303] M. Ahearn, M. Schaeffler, M. Vigeant, and R. D. Celmer, "The Just Noticeable Difference in the Clarity Index for Music, C80," *University of Hartford Acoustics Report EAC-2009-11*, 2009.

[304] J. S. Bradley, R. Reich, and S. G. Norcross, "A just noticeable difference in C50 for speech," *Applied Acoustics, 58 (2) 99-108*, 1999.

[305] A. H. Marshall, "A note on the importance of room cross-section in concert halls," *Journal of Sound and Vibration 5, 100-112*, 1967.

[306] A. Walther and C. Faller, "Interaural correlation discrimination from diffuse field reference correlations," *J. Acoust. Soc. Am. 133, 1496-1502*, 2013.

[307] N. Durlach, K. Gabriel, H. Colburn, and C. Trahiotis, "Interaural correlation discrimination: Ii. relation to binaural unmasking," *J. Acoust. Soc. Am. 79, 1548-1557*, 1986.

[308] K. Gabriel and H. Colburn, "Interaural discrimination: I. bandwidth and level dependence," *J. Acoust. Soc. Am. 69, 1394-1401*, 1981.

[309] M. Morimoto and K. Iidia, "Appropriate frequency bandwidth in measuring interaural cross-correlation as a physical measure of auditory source width." *Acoust. Sci. & Tech. 26, 2, 179-184*, 2005.

[310] C. Hak and R. Wenmaekers, "Measuring room acoustic parameters using a head and torso simulator instead of an omnidirectional microphone," in *38th International Congress and Exposition on Noise Control Engineering (Inter-Noise), Ottawa, Canada*, 2009.

Own Literature

[311] Neidhardt, Schneiderwind, and Klein, "Perceptual matching of room acoustics for auditory augmented reality in small rooms - a review," *Trends in Hearing, Vol. 22*, 2022. [Online]. Available: https://doi.org/10.1177/23312165221092919

[145] K. Brandenburg, F. Klein, A. Neidhardt, U. Sloma, and S. Werner, "Creating auditory illusions with binaural technology," in *The Technology of Binaural Understanding*. Springer Int. Publishing, 2020, pp. 623–663. [Online]. Available: https://doi.org/10.1007/978-3-030-00386-9_21

[313] K. Brandenburg, E. Cano Ceron, F. Klein, T. Köllmer, H. Lukashevich, A. Neidhardt, J. Nowak, U. Sloma, and S. Werner, "Personalized Auditory Reality," in *44th Annual Meeting on Acoustics (DAGA), Garching (Munich), Germany*, 2018.

[314] A. Neidhardt, "Effect of impaired early reflection patterns on plausibility and similarity of position-dynamic binaural AR audio," *unpublished.*, 2023.

[275] A. Neidhardt and A. M. Zerlik, "The availability of a real hidden reference affects the plausibility of position-dynamic auditory AR," *Frontiers in Virtual Reality*, 2021. [Online]. Available: https://doi.org/10.3389/frvir.2021.678875

[316] L. Remaggi, K. Hansung, A. Neidhardt, A. Hilton, and P. B. Jackson, "Perceived quality and spatial impression of room reverberation in VR reproduction from measured images and acoustics," in *23rd Int. Congress on Acoustics, Aachen, Germany*, 2019.

[317] A. Neidhardt, F. Klein, N. Knoop, and T. Köllmer, "Flexible python tool for dynamic binaural synthesis applications," in *142nd AES Convention, e-Brief 346, Berlin, Germany*, 2017.

[318] A. Neidhardt, "Data set: BRIRs for position-dynamic binaural synthesis measured in two rooms," in *5th Int. Conference on Spatial Audio, Ilmenau, Germany*, 2019.

[319] A. Neidhardt, A. M. Zerlik, and S. Kamandi, "BRIR data set for interactive listener translation in two rooms," Data set (1.0), Zenodo, Tech. Rep., 2020. [Online]. Available: zenodo.org/record/3457782

[278] C. Schneiderwind, A. Neidhardt, and D. Meyer, "Comparing the effect of different open headphone models on the perception of a real sound source," in *150th AES Convention, paper 10489, Online*, May 2021.

[321] A. Neidhardt and N. Knoop, "Binaural walk-through scenarios with actual self-walking using an HTC Vive," in *43rd Annual Conference on Acoustics, Kiel, Germany*, 2017.

[81] A. Neidhardt, A. Ignatious-Tommy, and A. D. Pereppadan, "Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets," in *144th AES Convention, paper 9987, Milan, Italy*, May 2018.

[323] A. Neidhardt, "Perception of the reverberation captured in a real room, depending on position and direction," in *22nd Int. Congress on Acoustics (ICA), Buenos Aires, Argentina*, 2016.

[324] A. Neidhardt and S. Kamandi, "Plausibility of an approaching motion towards a virtual sound source II: In a reverberant seminar room," in *152nd AES Convention, paper 10608, The Hague, The Netherlands / Online*, 2022.

[191] S. Werner, F. Klein, A. Neidhardt, U. Sloma, C. Schneiderwind, and K. Brandenburg, "Creation of Auditory Augmented Reality using a position-dynamic binaural synthesis system—Technical components, psychoacoustic needs, and perceptual evaluation," *Applied Sciences 11(3),*

Jan., p. 1150, 2021. [Online]. Available: https://doi.org/10.3390/app11031150

[326] A. Neidhardt and B. Reif, "Minimum BRIR grid resolution for interactive position changes in dynamic binaural synthesis," in *148th AES Convention, paper 10371, Vienna, Austria*, 2020.

[327] A. Neidhardt and C. Schneiderwind, "The influence of the DRR on audiovisual coherence of a real loudspeaker playing virtually over headphones," in *47th German Annual Conference on Acoustics, Vienna, Austria.*, 2021.

[328] C. Schneiderwind and A. Neidhardt, "Discriminability of concurrent virtual and real sound sources in an augmented audio scenario," in *152nd AES Convention, paper 10604, The Hague, The Netherlands / Online*, 2022.

[329] R. Gupta, J. He, R. Ranjan, W.-S. Gan, F. Klein, C. Schneiderwind, A. Neidhardt, K. Brandenburg, and V. Välimäki, "Augmented/Mixed Reality Audio for Hearables: Sensing, control, and rendering," *IEEE Signal Processing Magazine*, 2022.

[130] C. Schneiderwind and A. Neidhardt, "Perceptual differences of position dependent room acoustics in a small conference room," in *Int. Symposium on Room Acoustics (ISRA), Amsterdam, Netherlands*, 2019.

[146] F. Klein, S. Werner, and T. Mayenfels, "Influences of training on externalization of binaural synthesis in situations of room divergence," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 178–187, 2017.

[188] A. Neidhardt, "Distance perception in virtual auditory environments with a moving avatar," in *42nd Annual Conference on Acoustics, Aachen, Germany*, 2016.

[287] ——, "Data set and physical analysis: BRIRs and SRIRs for walking toward, past and behind virtual loudspeakers in two rooms," in *154th AES Convention, Espoo/Helsinki, Finland*, 2023.

[334] ——, "Data set of measured room impulse responses: BRIRs, RIRs and SRIRs for position-dynamic binaural auralization in two rooms," Technische Universität Ilmenau, Tech. Rep., 2023. [Online]. Available: zenodo.org/record/7838178

[335] A. Neidhardt and C. Schneiderwind, "Working Sheet: Applied and Virtual Acoustics Seminar - Room Acoustic Parameters," 2021, Institute of Media Technology, Technische Universität Ilmenau, Germany.

Supervised Student Projects

The lists are sorted by their relevance for this PhD thesis.

Master Theses:

- Afrooz Nasrollahnejad, "Evaluating plausibility of interactive virtual acoustic environments for headphone reproduction", 2019, TU Ilmenau.
- Samaneh Kamandi, "Perception of simplification of room acoustics in dynamic binaural synthesis for listener translation", 2019, TU Ilmenau.
- Christian Schneiderwind, "Analyse von Raumakustik in kleinen Räumen mit Hilfe von Eigenmike-Aufnahmen" [Analysis of the acoustic in small rooms using Eigenmike recordings], 2018, TU Ilmenau.
- Alexandra Wahl, "Automatische Zuordnung ähnlich wahrgenommener Räume" [Automatic assignment of perceptually similar rooms], TU Ilmenau, 2020.
- Aravindan Benjamin, "Psychoacoustic evaluation of binaural auralization of virtual room acoustics", 2017, TU Ilmenau.
- Anson Davis Pereppadan, "Plausibility of binaural walk-through-scenarios considering different simplifications", 2018, TU Ilmenau.
- Nils Merten, "Gegenüberstellung verschiedener Algorithmen zur Erzeugung einer positionsdynamischen Binauralisation aus stark reduzierten Messdaten" [Comparison of different algorithms for creating position-dynamic binaural auralizations from very sparse measured data], 2019, TU Ilmenau.
- Anna Maria Zerlik, "Untersuchung der Authentizität von Binaural Walk-Throughs" [Investigating the authenticity of binaural walk-through scenarios], 2019, TU Ilmenau.
- Alexandra Draghici, "Virtual sound source behavior in the real environment", TU Ilmenau, 2019.
- Avinash Tarale, "Perception of Simplified Representations of a Wall in a Virtual Acoustic Environment", 2018, TU Ilmenau.
- Tarek Al Sibai, "Quick online adaptation of the late reverberation of a binaural reproduction to the real room", TU Ilmenau, 2020.
- Nawres Haitham Abdulqader Al-Janabi, "Interactive exploration of a virtual auditory environment using self-created sounds", 2017 TU Ilmenau
- Bibek Chandra Bhattarai, "Binaural auralization of virtual room acoustics", 2017 TU Ilmenau
- Michael Blau, "Implementierungsmöglichkeiten für Wiedergabesysteme mit dynamischer Binauralsynthese" [Options for implementing reproduction systems based on binaural synthesis], TU Ilmenau, 2018.

- Niklas Knoop, "Orientierung in einem virtuellen Raum mit beweglichem Avatar" [Orientation in a virtual room with a moving avatar], 2016 TU Ilmenau
- Thomas Thron, "Raumakustische Simulation aus Basis geometrischer und optischer Raumparameter" [Room acoustic simulation based geometrical and optical room parameters], TU Ilmenau, 2015.
- Tahereh Afghah, "Perception/ Evaluation of distance after a continuous change of distance", TU Ilmenau, 2015.
- Manan Lamba, "Contribution of listener translation on the perceived externalization in binaural reproduction", 2019, TU Ilmenau.

Bachelor Theses:

- Nils Merten, "Distanzwahrnehmung in virtuellen auditiven Umgebungen mit dynamischem Avatar" [Distance perception in virtual auditory environments with a dynamic avatar], TU Ilmenau, 2017.
- Dominik Meyer, "Einfluss des Kopfhörermodells auf reale und virtuelle Schallquellen in Augmented Acoustics Anwendungen" [Influence of the headphones model on real and virtual sound sources in augmented acoustic applications], TU Ilmenau, 2020.
- Oliver Doll, "Extraktion akustischer Umgebungsmerkmale aus binauralen Audiostreams" [Extraction of acoustic properties of the environment from binaural audio streams], TU Ilmenau, 2020.
- Marius Seipel, "Entwicklung und Evaluierung von Messmethoden zur In-Situ Bestimmung der Raumakustik und der Gewinnung von binauralen Raumimpulsantworten" [Development and evaluation of methods for in-situ estimation of the acoustics and binaural room impulse responses of a room], TU Ilmenau, 2019.
- Stefan Fichna, "Perception of room acoustical differences between listening positions in the case of low direct sound", Jade-Hochschule Oldenburg, 2018.
- Tatiana Surdu, "Investigations of the perceptual mixing time for dynamic binaural synthesis in small rooms for interactive position changes", TU Ilmenau, 2019.
- Anna Maria Zerlik, "Einfluss selbst ausgelöster Geräusche bei der Orientierung in einer virtuellen auditiven Szene" [Impact of self-generated sounds on the orientation within a virtual auditory scene], 2015, TU Ilmenau.
- Tobias Boley, "Identifikation relevanter Parameter zur Orientierung in einer virtuellen auditiven Szene" [Identification of relevant parameters for the orientation within a virtual auditory scene], TU Ilmenau, 2015.
- Georg Rolapp, "Wahrnehmung der eigenen Hörposition im Raum nach einer Bewegung" [Perception of the own listening position in a room after moving], TU Ilmenau, 2016.
- Niklas Knoop, "Der Einfluss visueller Eindrücke auf die auditive Raumwahrnehmung" [Influence of visual impressions on the auditory room perception], TU Ilmenau, 2015.

Media Projects:

- Afrooz Nasrollahnejad und Kai Peter Jurgeit, "Psychoacoustic Evaluation of Algorithms for Continuously Changing Listening Positions in Interactive VAE", 2017, TU Ilmenau.
- Robin Ritter und Christian Schneiderwind, "Physikalische und wahrnehmungsbezogene Analyse der Raumakustik an verschiedenen Positionen im Raum" [Physical and perceptual analysis of the acoustics at different positions in a room], 2017, TU Ilmenau.
- Manan Lamba, Chenyao Diao und Shuang Wang, "Investigating the perception of a nearby wall by exploring the virtual room with self-produced oral sounds", 2018, TU Ilmenau.
- Alby Ignatious Tommy und Anson Davis Pereppadan, "Walking towards a virtual loudspeaker - a realization", 2017, TU Ilmenau.
- Bernhard Fiedler and Tobias Heinl, "Einfluss der Simulationsparameter auf die Wahrnehmung virtueller Räume für bewegliche Hörer" [Influence of simulation parameters on the perception of virtual room for moving listeners], TU Ilmenau, 2015.
- Kenneth Warmuth, Oliver Doll, Josephin Wiesner, "Automatic estimation of room acoustic parameters from binaural microphone signals", TU Ilmenau, 2021.
- Nuzhat Gawhar and Sadia Ferdous Snigdha, "Audiovisual matching of room acoustics", TU Ilmenau, 2021.

List of Figures

List of Tables

A. Detailed analysis of room acoustics in both rooms based on measured data

This appendix contains several text passages from the conference paper "Data set data and physical analysis: BRIRs and SRIRs for walking toward, past and behind virtual loudspeakers in two rooms" that I published after the submission of this thesis in 2023 [287].

Room size, volume, Schroeder frequency

The seminar room has a size of 9.9 m×4.7 m×3.1 m, a volume of V = 144.3 m^3 and a reverberation time of T_{30} = 1.1 s (mean of 500 Hz and 1 kHz). The Schroeder frequency [160] can be determined by

$$f_{Schroeder} = 2000\sqrt{\frac{T}{V}} \tag{A.1}$$

The Schroeder frequency of the seminar room is 165 Hz.

The listening lab has a size of 8.4 m×7.6 m×2.8 m, a volume of V = 179 m^3 and a reverberation time T_{30} = 0.23 s. Hence, the Schroeder frequency is 72 Hz.

However, Skålevik [161] suggests that in small rooms in practice, the transition range between modal behavior and the typical properties of high frequencies is at lower frequencies than the Schroeder frequency.

Reverberation Time

Tab. A.1 gives an overview of the estimated frequency-dependent reverberation times T_{20} and T_{30} in the octave bands for listening lab (LL) and seminar room (SR). The values have been determined based on the average of the RIRs measured with the centre microphone of the array in each room.

Table A.1.: Reverberation times T_{20} and T_{30}

Freq	T_{20} LL	T_{30} LL	T_{20} SR	T_{30} SR
125 Hz	257 ms	250 ms	1260 ms	1212 ms
250 Hz	229 ms	239 ms	958 ms	997 ms
500 Hz	238 ms	241 ms	1020 ms	1052 ms
1 kHz	223 ms	225 ms	1172 ms	1178 ms
2 kHz	201 ms	209 ms	1171 ms	1167 ms
4 kHz	188 ms	189 ms	934 ms	945 ms

Critical distance and DRR

The critical distance or reverberation radius is defined as the distance from a sound source in the room at which the level of direct sound, in the process of its wavefront spreading, becomes equal to

a reflected sound level. Textbooks usually present the following equation to determine the critical distance [10]:

$$r_c = 0.057\sqrt{\frac{\gamma V}{T}} \tag{A.2}$$

In small rooms, the actual critical distance is often considerably different from the value determined with the well-known equation [289]. Therefore, it should be estimated based on acoustic measurements. The direct-to-reverberant-energy (DRR) ratio is a suitable indicator. A DRR of 0 dB indicates that the energy of direct sound and reverberation are equal, larger DRR values indicate a dominance of the direct sound and negative values a dominance of the reverberation. Fig. A.1 shows the progress of the DRR along the nine positions at both ears of the Kemar and at the centre microphone of the array.
In the listening lab, the translation line is entirely located within the critical distance for the direct reproduction scenario. In contrast, for the indirect reproduction, all listening positions are located outside the critical distance.

In the seminar room, only a part of the translation line is located within the critical distance of the sound source. A listener walking along the given line will cross the critical distance. In a distance of about 2 m from the loudspeaker, which is roughly at the centre of the translation line, the direct and reverberant sound energy are equivalent in the BRIRs for 0°. For the indirect reproduction, again, all listening positions are located outside the critical distance. Fig. A.1 shows the DRR progress along the line in both rooms.

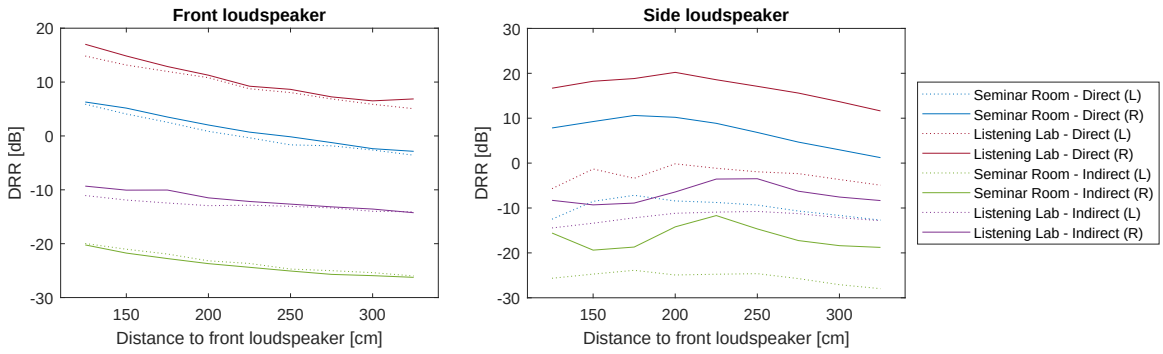


Figure A.1.: Direct-to-Reverberant-Energy Ratio for both loudspeakers along the 2 m-line with an azimuth angle of 0°

Predicted Perceptual Mixing Time

System-based prediction Lindau et al. [150] proposed system-based models to predict the average perceptual mixing time t_{mp50} to achieve a transparent simulation for 50% of the listeners and t_{mp95} for the majority, considering the room volume. For the **seminar room** with V = 144.2 m^3, the estimated perceptual mixing time would be

$$t_{mp95} = 0.0117\,V + 50.1\,ms = 51.79\,ms$$
$$t_{mp50} = 0.58\,\sqrt{V} + 21.2\,ms = 28.16\,ms.$$

The room contains some furniture, as shown in Fig. 4.21. Thus, the volume of interest may be smaller, but according to that model, the effect on the estimated perceptual mixing time is minimal in small rooms.
For the **listening lab** with V = 179 m^3, the system-based mixing time estimates are

$$t_{mp95} = 0.0117\,V + 50.1\,ms = 52.19\,ms$$
$$t_{mp50} = 0.58\,\sqrt{V} + 21.2\,ms = 28.96\,ms.$$

Signal-based prediction In addition, Lindau et al. [150] proposed the prediction with signal-based models considering t_{mix_Abel} according to Abel and Huang [272]. AKtools [273] provide a script to estimate t_{mix_Abel} from a measured RIR. At each of the nine positions along the line, RIRs were measured for the four different arrangements of the sound source. The script was applied to all of these RIRs and for the **seminar room**, the resulting values ranged from 30 ms to 85 ms with a mean of $t_{mix\_Abel} = 49.4\,ms$. Consequently, t_{mp} can be predicted with

$$t_{mp95} = 1.8 \cdot t_{mix\_Abel} - 38\,ms = 50.9\,ms$$
$$t_{mp50} = 0.8 \cdot t_{mix\_Abel} - \;\;8\,ms = 31.5\,ms.$$

For the **listening lab**, the estimated t_{mix_Abel} varied from 17 ms to 42 ms for the different listening positions with a mean of 20.6 ms. This leads to following predictions of t_{mp} according to the proposed model:

$$t_{mp95} = 1.8 \cdot t_{mix\_Abel} - 38\,ms = -0.9\,ms$$
$$t_{mp50} = 0.8 \cdot t_{mix\_Abel} - \;\;8\,ms = 8.5\,ms.$$

The value indicates that a prediction of a mixing time in the room is difficult, probably because a considerable diffuse reverberation is not established in this rather dry room.

Physical analysis of early reflections

Listening Lab The perceptual effects of strong ER highly depend on the delay, level and angle of incidence relative to the direct sound. Tab. A.2 gives an overview of the six most prominent reflections at both ends of the translation line for the four source scenarios. Delay and level of the reflections with respect to the direct sound were extracted with the *paraspax*-toolbox according to Arend et al. [85], as well as the time-of-arrival (TOA) after the start of the sound emission, the level relative to the global maximum of the scene, and the direction of incidence of the single reflections. The energy of the direct sound arriving at the different listening positions varies substantially and with it, the relative level of the reflections.

Table A.2.: Delay and energy of early reflections in the listening laboratory

Delay[ms]	L_{dir}[dB]	TOA[ms]	L_{glob}[ms]	DOA_{az}[°]	DOA_{el}[°]	Delay[ms]	L_{dir}[ms]	TOA[ms]	L_{glob}[ms]	DOA_{az}[°]	DOA_{el}[°]
Pos1 (125cm)	Front Dir	3.27	0.00	0.00	0.00	Pos1 (125cm)	Side Dir	3.50	-1.38	-113.96	0.00
6.71	-20.94	10.04	-30.17	-9.36	-68.51	6.46	-17.57	10.04	-29.58	-115.30	-64.89
4.27	-21.12	7.60	-30.35	1.27	64.26	13.19	-17.97	16.77	-29.98	96.62	-0.58
1.73	-21.18	5.06	-30.41	-7.31	-18.13	1.54	-19.60	5.13	-31.61	-109.10	-10.19
21.75	-23.31	25.08	-32.55	179.18	-1.76	17.52	-20.41	21.10	-32.42	96.87	-0.31
24.02	-24.04	27.35	-33.28	-145.26	-41.34	16.65	-20.94	20.23	-32.95	96.03	-27.53
28.50	-26.48	31.83	-35.71	-4.80	5.67	4.13	-20.96	7.71	-32.97	-112.83	62.93
Pos9 (325cm)	Front Dir	9.04	-9.09	0.00	0.00	Pos9 (325cm)	Side Dir	5.40	-10.50	-36.87	0.00
2.35	-11.17	11.48	-29.20	-2.73	20.45	5.44	-10.53	10.85	-30.20	-40.08	-59.94
4.02	-11.35	13.15	-29.38	-2.04	-45.56	16.73	-14.43	22.15	-34.10	79.20	0.49
10.21	-13.27	19.33	-31.30	179.29	-3.15	3.33	-14.87	8.75	-34.54	-37.09	53.19
11.35	-14.11	20.48	-32.14	-163.84	-17.92	10.63	-15.16	16.04	-34.83	-170.39	10.18
12.58	-15.17	21.71	-33.20	-164.96	-25.31	14.19	-15.32	19.60	-35.00	140.66	7.94
7.19	-17.59	16.31	-35.62	57.13	0.53	17.65	-15.70	23.06	-35.37	80.14	14.89
Pos1 (125cm)	Front Indir	4.17	-21.65	0.00	0.00	Pos1 (125cm)	Side Indir	4.38	-21.08	-109.98	0.00
19.69	-0.15	23.92	-30.59	-1.16	11.34	18.38	5.49	22.81	-25.45	-93.83	-3.92
18.77	-0.37	23.00	-30.81	0.26	3.93	1.58	-2.02	6.02	-32.95	-138.38	-24.51
1.81	-2.30	6.04	-32.74	-18.04	-47.90	5.73	-4.35	10.17	-35.29	-115.26	-60.98
5.81	-3.24	10.04	-33.65	-13.41	-68.35	19.71	-4.85	24.15	-35.79	-93.14	12.75
21.29	-3.77	25.52	-34.21	-2.71	15.67	17.08	-5.71	21.52	-36.65	165.00	-24.11
22.94	-6.06	27.17	-36.50	-63.89	-7.96	22.10	-6.30	26.54	-37.23	-67.16	-1.43
Pos9 (325cm)	Front Indir	10.17	-28.16	0.00	0.00	Pos9 (325cm)	Side Indir	6.13	-23.96	-42.51	0.00
18.83	2.85	28.85	-34.69	-31.30	7.38	16.92	-1.26	22.85	-34.91	-80.78	7.33
19.63	0.68	29.65	-36.86	-0.94	8.59	5.00	-2.37	10.94	-36.01	-57.01	-58.44
21.19	-2.23	31.21	-39.76	-5.85	-0.96	15.73	-2.70	21.67	-36.34	-79.04	5.99
14.48	-2.28	24.50	-39.81	-13.67	12.35	1.08	-3.60	7.02	-37.24	-109.20	-37.80
3.65	-2.37	13.67	-39.91	-3.21	-46.89	13.00	-4.35	18.94	-37.99	-91.01	-30.13
31.10	-2.49	41.13	-40.02	-19.88	-23.73	2.13	-4.57	8.06	-38.21	-96.98	-39.37

Seminar room Tab. A.3 gives an overview of the six most prominent reflections at both ends of the translation line for the four source scenarios.

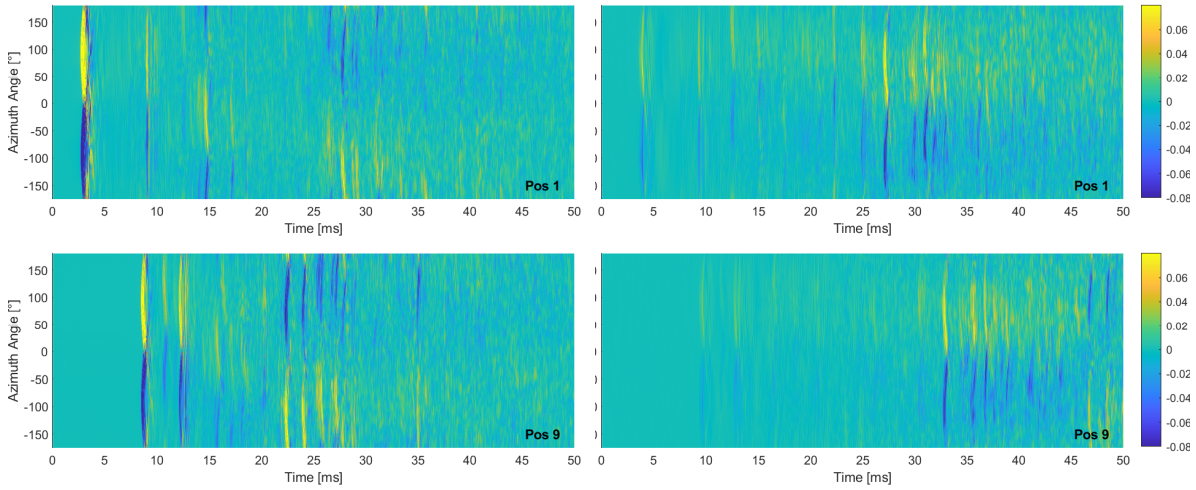


Figure A.2.: Direction dependent BRIRs in the time domain: |L|-|R| at positions Pos1 and Pos9 for the cases Front Direct and Front Indirect.

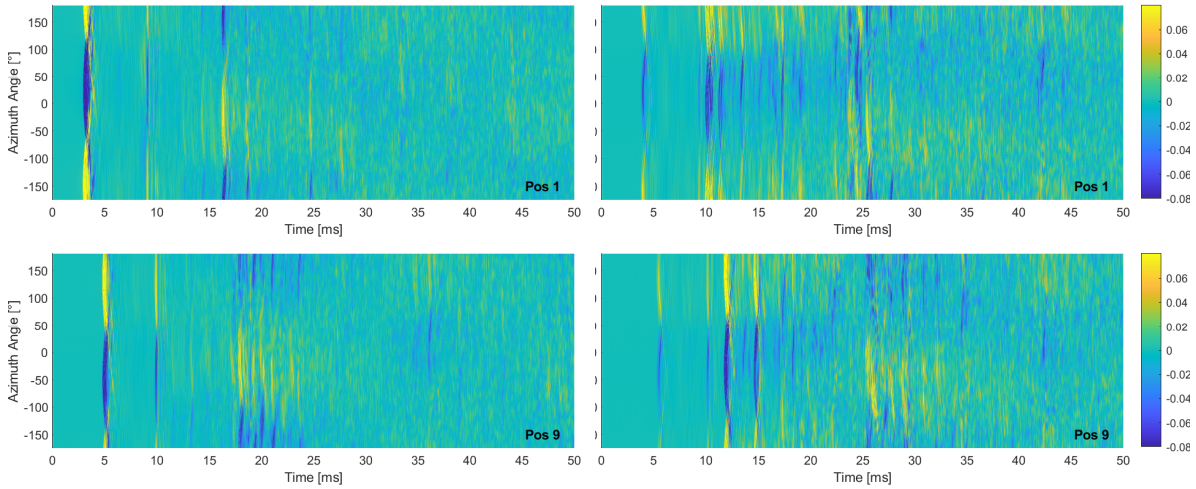


Figure A.3.: Direction dependent BRIRs in the time domain: |L|-|R| at positions Pos1 and Pos9 for for the cases Side Direct and Side Indirect.

In the Front-Direct scenario, the six strongest reflections arrived at the closest listening position (Pos1) about 6 to 28 ms after the direct sound with levels from -20.8 to -15.3 dB relative to the direct sound and at the farthest position (Pos9) 4 to 20 ms after the direct sound with levels of -10 to -5 dB. For the Front-Indirect scenario, the reflections arrived between 23 to 40 ms with levels of 0.65 to 7.4 dB. For the Side-direct case, the six strongest reflections arrive at Pos1 again about 6 to 28 ms after the direct sound with a level of -19.6 to -12.8 dB, and at Pos9 6 to 36 ms after the direct sound with levels of -11.4 to -5.7 dB. For the Side-Indirect case, the six strongest reflections arrived within a similar time range as for the direct case, but with 0.7 to 9.2 dB at Pos1 and 3.8 to 13.6 dB were again stronger than the direct sound. Tab. A.3 provides a detailed overview.

Table A.3.: Delay and energy of early reflections in the seminar room

Delay[ms]	L_{dir}[dB]	TOA[ms]	L_{glob}[ms]	DOA_{az}[°]	DOA_{el}[°]	Delay[ms]	L_{dir} [ms]	TOA[ms]	L_{glob}[ms]	DOA_{az}[°]	DOA_{el}[°]
Pos1 (125cm)	Front Dir	3.27	0.00	0.00	0.00	Pos1 (125cm)	Side Dir	3.50	-1.38	-113.96	0.00
6.13	-15.28	9.40	-15.28	-1.32	69.70	13.33	-12.76	16.83	-14.14	112.59	-0.51
25.02	-17.14	28.29	-17.14	-169.04	0.96	5.94	-13.49	9.44	-14.87	-116.68	67.71
24.19	-17.46	27.46	-17.46	-161.13	-33.53	6.75	-14.72	10.25	-16.10	-123.29	-26.12
11.83	-19.10	15.10	-19.10	91.48	-1.08	12.38	-14.89	15.88	-16.27	161.37	-36.88
25.98	-20.13	29.25	-20.13	168.98	-8.82	15.54	-17.26	19.04	-18.64	113.92	30.20
28.19	-20.80	31.46	-20.80	154.54	-18.47	27.63	-19.63	31.13	-21.01	-109.70	4.67
Pos9 (325cm)	Front Dir	9.04	-9.09	0.00	0.00	Pos9 (325cm)	Side Dir	5.40	-10.50	-36.87	0.00
4.15	-5.16	13.19	-16.09	-5.11	-47.67	12.98	-5.69	18.38	-16.19	107.15	-26.88
13.81	-5.35	22.85	-16.28	-162.72	-1.34	14.38	-6.19	19.77	-16.69	131.60	-2.31
12.81	-7.34	21.85	-18.26	-117.98	4.80	5.56	-6.77	10.96	-17.27	-52.89	-36.11
18.17	-9.35	27.21	-20.27	-91.77	-58.54	17.44	-9.74	22.83	-20.25	107.42	-64.67
15.42	-9.67	24.46	-20.59	-158.64	49.57	16.04	-9.91	21.44	-20.42	135.14	16.22
19.54	-9.97	28.58	-20.90	160.32	0.92	35.73	-11.43	41.13	-21.94	123.77	79.68
Pos1 (125cm)	Front Indir	4.17	-21.65	0.00	0.00	Pos1 (125cm)	Side Indir	4.38	-21.08	-109.98	0.00
26.31	2.45	30.48	-19.20	7.99	-23.58	6.10	9.22	10.48	-11.87	-89.08	1.13
23.40	2.06	27.56	-19.59	9.07	-0.96	19.90	4.12	24.27	-16.98	110.83	-0.96
22.58	1.84	26.75	-19.82	78.75	1.49	7.33	3.60	11.71	-17.50	-90.40	-24.29
28.73	1.65	32.90	-20.00	1.03	-14.02	21.52	3.20	25.90	-17.89	111.95	16.84
27.25	1.13	31.42	-20.52	-12.27	-0.40	9.31	1.18	13.69	-19.91	-88.80	25.39
35.54	0.65	39.71	-21.01	14.95	44.76	23.77	0.70	28.15	-20.39	118.77	6.62
Pos9 (325cm)	Front Indir	10.17	-28.16	0.00	0.00	Pos9 (325cm)	Side Indir	6.13	-23.96	-42.51	0.00
40.25	7.41	50.42	-20.72	122.02	-58.82	6.15	13.57	12.27	-10.40	-62.12	0.25
23.04	6.99	33.21	-21.15	6.90	-0.33	8.96	9.15	15.08	-14.82	-62.17	34.70
39.21	6.80	49.38	-21.34	-117.60	-43.00	20.19	8.44	26.31	-15.53	119.04	-1.03
27.02	6.51	37.19	-21.62	-9.43	0.56	23.06	7.22	29.19	-16.74	137.32	5.28
25.71	5.82	35.88	-22.32	8.66	-13.29	21.29	5.60	27.42	-18.37	133.63	-38.25
28.17	5.31	38.33	-22.83	-23.41	-49.11	25.21	3.75	31.33	-20.22	124.22	-44.76

Predicted saliency of early reflections

Brinkmann et al. [80] proposed a model for predicting the saliency of the individual early reflections. From the psychoacoustic experiment they conducted to verify their model, they conclude that it may be sufficient to auralize the six strongest early reflections. With the help of their toolbox, the saliency of early reflections was analyzed for the SRIRs measured at the first and last position of the translation line in both the listening lab and the seminar room.

Fig. A.4 visualizes the predictions of perceptually salient reflections according to Brinkmann et al. [80] in the listening laboratory. Close to the front of the loudspeaker (graph (A)), only a single early reflection seems to be relevant. Two meters further away (graph (B)), a few more reflections have to be taken into account according to the model. Turning the loudspeaker by 180° (graph (C)) leads to a further substantial increase in the number of salient reflections. For the loudspeaker on the side of the translation line, similar relations were observed.

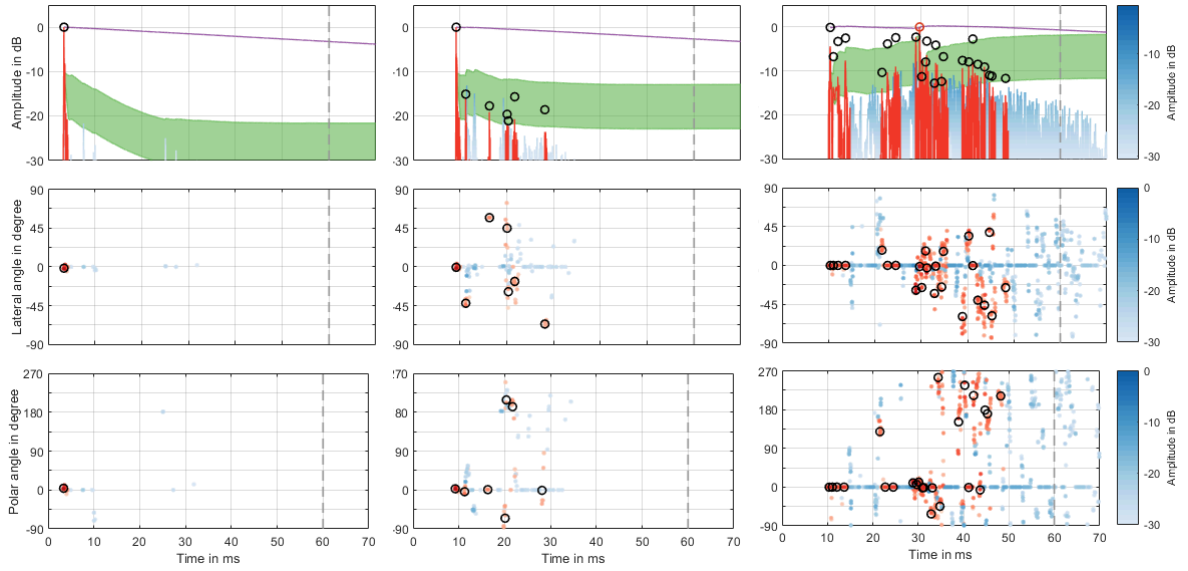


Figure A.4.: Salient early reflections predicted according to Brinkmann et al. [80] for three selected positions in the **listening lab**: (A) Pos1 in Front-Direct scenario - listening position in front of the loudspeaker, (B) Pos9 in Front-Direct scenario - most distant listening position, (C) Pos9 in Front-Indirect scenario - most distant listening position with loudspeaker turned around.

Fig. A.5 shows the predictions of perceptually salient reflections. The observations are similar to the listening lab. Close to the front of the loudspeaker (graph (A)), only very few - approximately six - early reflections are predicted as salient. Two meters further away (graph (B)), already many more reflections are perceptually relevant according to the model. Turning the loudspeaker by 180° (graph (C)) leads to a further substantial increase in the number of salient reflections. For the loudspeaker on the side of the translation line, similar relations were observed.

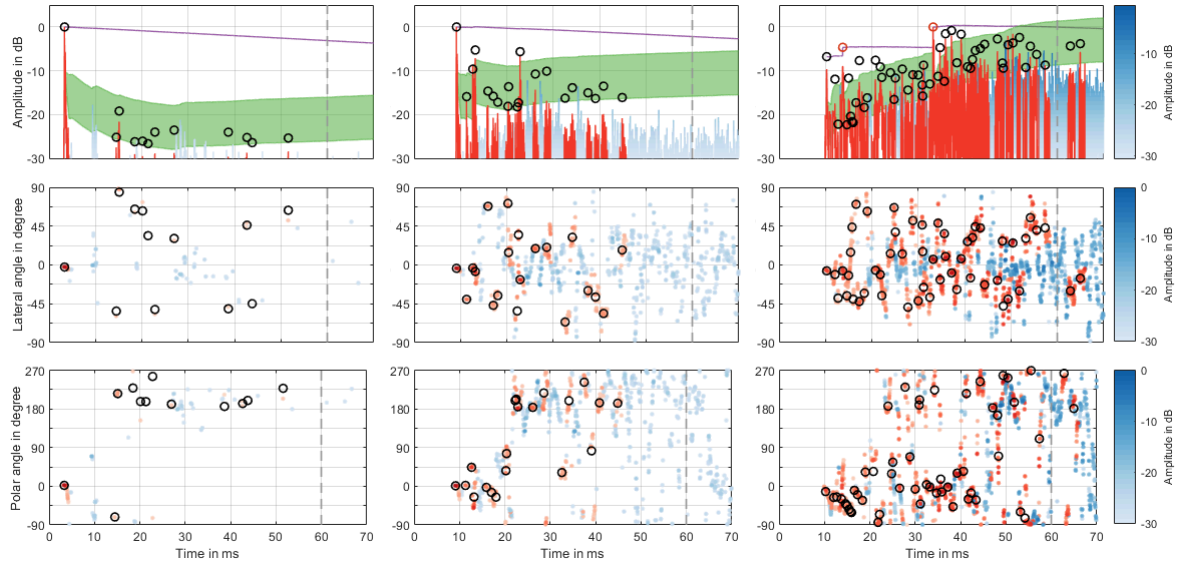


Figure A.5.: Salient early reflections predicted according to Brinkmann et al. [80] are visualized in red for three selected positions in the **seminar room**: (A) Pos1 in Front-Direct scenario - listening position in front of the loudspeaker, (B) Pos9 in Front-Direct scenario - more distant listening position, (C) Pos9 in Front-Indirect scenario - more distant listening position with loudspeaker turned around. The green area marks the modeled masking threshold function, the purple line visualizes the echo threshold and the gray dashed line indicates the predicted perceptual mixing time of approximately 50 ms after the direct sound.

A.1. Summary

For this thesis, a data set of acoustic impulse responses (BRIRs, RIRs and SRIRs) was measured in two rooms to investigate the perceptual effects of simplified acoustic room representations. The two rooms exhibit reverberation times of about $T_{30} = 0.23\,s$ (listening laboratory) and about $T_{30} = 1.1\,s$ (seminar room).

The data set is available online [334]:
zenodo.org/record/7838178

This appendix presents a physical analysis of the measure data, including a frequency-dependent analysis of the reverberation time, the direct-to-reverberant-energy ratio, and further room acoustic parameters. Moreover, a prediction of the perceptual mixing time based on physical properties is derived.

The availability of BRIRs and SRIRs measured at the same listening positions for different positions and orientations of the sound source in two different reverberation settings, makes the data set very interesting for physical and perceptual evaluations of auralization methods based on interpolation, extrapolation or parametric approaches. Particularly the cases with loudspeakers facing away from the listening positions are critical test cases that could reveal weaknesses of immature auralization approaches.

B. Room acoustic parameters and their correspondence to perception

This appendix summarizes parameters that are relevant for the physical and perceptual analysis of the (small) rooms considered in this thesis. On the one hand, this includes several of the parameters recommended by the standard in the context of concert hall acoustics. On the other hand, further parameters were added which turned out to be useful in the discussion and analysis of small perceptual differences in the room acoustics. The overview has already been published informally as a working sheet for students in a seminar on room acoustics with in the lecture "Applied and Virtual Acoustics" at the University in Ilmenau [335].

Reverberation time (RT_{60}, RT_{30}, RT_{20}) The *reverberation time (RT60)* of a room is defined as duration until the spatially averaged sound energy decreased by 60 dB after switching off the sound source [51]. It can be determined from a measured impulse response using the *Energy Decay Curve (EDC)* proposed by Schroeder [290]. It is calculated by backwards integration (also: tail integration) of the squared impulse response.

$$EDC(t) = \int_t^\infty h(\tau) d\tau$$

In the optimal case of noise-free room impulse response, the time interval for integration could be $[0, \infty]$ as suggested in the standard. Realistic measurement data usually contains a noise floor of a certain level which will influence the EDC. To minimize this influence, the point of time when the decaying sound energy is as low as the energy level of the noise floor, can be used as the starting point of the backwards integration to avoid a wrong increase of the estimated reverberation times. RT60 is the duration starting from the -5 dB decay of the sound energy in the room to a decay of -65 dB. Therefore, a sufficient Signal-to-Noise-Ratio (SNR), the Peak-to-Noise-Ratio (PNR) or Impulse-response-to-Noise-Ratio (INR) play an important role. If the SNR is too low, the -65 db decay of the reverberation may be below the level of the noise floor and RT60 cannot be determined adequately. This is of particular interest with regard to a determination in the separate frequency bands. Sometimes, the achievable SNR is also limited by practical constraints.

The reverberation time can be determined based on a smaller dynamic range and then be extrapolated for a decay of 60 dB. If it is calculated from a decay range of -5 dB to -25 dB the value is labeled as RT_{20} or RT_{30} for a range of 5 dB to 35 dB. According to the standard [51] this procedure is repeated for each octave band separately. For the determination of the reverberation time, the decay times estimated for the 500 Hz and 1000 Hz octave band are averaged.

Furthermore, the standard [51] demands to measure room impulse responses with an omni-directional sound source and an omni-directional microphone at various positions in the room and calculate the average. This standard has been developed for performance spaces like concert halls. DIN EN ISO 3382-2 [9] describes the standardized procedure to determine the reverberation time in ordinary rooms like class rooms or living rooms and distinguishes three different degrees of precision, namely *short*, *standard*, and *precise* measurement. For short and standard measurements, an omni-directional sound source is not strictly required. A standard determination of RT also demands for measurements for at least X different combinations of source and receiver positions. In

smaller rooms, achieving a suitable dynamic range can be challenging. A dynamic range of 35 dB is necessary to get a 20 dB analysis range to determine RT20.

It is also worth noting that RT values can vary with the implemented algorithm [291]. For impulse responses with an exponential decay and a steady noise floor or without noise, the deviations are minimal. However, in small rooms, a perfectly exponential decay cannot be assumed. ISO 3382-2 [9] proposed to calculate the degree of curvature to check the validity of the estimated reverberation times. It can be calculated with $C = 100 * (\frac{RT30}{RT20} - 1)$. C should be within the range of 0-5 %.

Early Decay Time (EDT) According to [292], in concert hall acoustics the Early Decay Time (EDT) relates better to the perception of reverberance than the reverberation time T_{60}. The procedure to determine this value is similar to that of RT, only by using decay from 0 dB to -10 dB and extrapolating it linearly to -60 dB.

According to DIN EN ISO 3382-1 [51], the JND of the EDT is a relative change of 5 %. In an investigation with SRIRs of three different concert halls, Del Solar Dorrego and Vigeant [293, 95] observed JNDs of about 18-20 % for broad band conditions.

Critical distance in real rooms

The critical distance, also called reverberation radius, is the distance from a source at which the level of the direct sound equals the level of the reflected sound. Based on a statistical model, the following equation was proposed to determine an approximation of the critical distance.

$$r_{crit} = 0.057\sqrt{\frac{\gamma V}{T}} \tag{B.1}$$

r_{crit} is the critical distance in m, T is the reverberation time in s, V is the volume of the room in m^3 and γ is the source directivity [294].

This equation is based on a statistical model that assumes homogeneous and diffuse reverberant sound field, wherein the sound energy does not depend on position or direction. Particularly in rooms with high absorption, the sound field does not have these properties. The critical distance can differ substantially from the theoretical value provided by eq. B.1 [289]. Mijić and Mašović propose to determine the critical distance by measuring acoustic impulse responses at different positions with varying distances to the source, estimated the direct and the reverberant energy in these impulse responses and fit a corresponding curve into this data. In their study with four different small rooms, the reverberation radius was determined with measurements. The well-known theoretical equation caused errors of up to $\pm 50\,\%$ in these cases. In contrast, the measurement based method showed an error of $\pm 10\,\%$.

Aura-Puchades and Berardi [295] proposed a new formula to calculate the critical distance in rooms with non-diffuse sound fields, for example, rooms with non-uniformly distributed sound absorption. It considers the direction-dependent absorption A_i and thus, depends on the direction as well.

$$r_{critD_i} = \sqrt{\frac{\gamma A_i}{16\pi}} \tag{B.2}$$

In real rooms, the determination of A_i is not straight-forward and will require a complex capturing procedure of the sound field. Moreover, in real rooms, the critical distance is likely to vary with the position and orientation of the source due to the variations in the non-diffuse reverberation. To get an overview of the relations between direct and reverberant sound energy, it is of interest to analyze the direct-to-reverberant energy ratio of available measurement data.

Direct-to-reverberant energy ratio (DRR) The Direct-to-reverberant energy ratio (DRR) is the energy ratio of direct and reverberant sound energy at a certain position in the room. It plays an essential role in the auditory perception of the egocentric distance to a sound source in rooms [296][2] [297]. The DRR can be determined from omni-directionally measured room impulse responses by separating direct and reverberant sound energy and calculating the DRR according to eq. B.3.

$$DRR = 10log_{10} \frac{\int_{0ms}^{T} h^2(t)dt}{\int_{T}^{\infty} h^2(t)dt} \tag{B.3}$$

with $h(t)$ being the measured impulse response and T being the duration of the direct sound. T is often set to 2 ms ($\pm 1\,ms$ from the peak).

The direct sound energy varies considerably with the distance to the sound source. In case of a directional sound source, it also varies with the relative direction to the sound source [136, 137]. Thus, the DRR shows strong deviations between different positions in a room. However, also the reverberant energy can vary substantially between the positions of a room.

Zahorik [253] examined the sensitivity to DRR changes within a single environment for ratios ranging from 0 dB to 20 dB. In this test setup, discrimination thresholds of of 5-6 dB were determined for all considered types of signals. Reichardt and Schmidt [298], in contrast, observed discrimination thresholds of about 2 dB. This deviation may be result of different manipulation approaches, for example, whether the direct sound energy or the reverberant sound energy is adjusted or whether the overall level is kept constant, which might require an adjustment of both. Larsen et al. [139] observed that the discrimination thresholds for DRR depend on the total amount of the DRR. For a DRR of -10 dB to 0 dB a difference of 2-3 dB could just be noticed, while at -20 dB or 10 dB JNDs were around 6-8 dB. Overall, the determination of JNDs for the DRR remains challenging and complex, since even cases with equal DRR values can sound very different. The temporal and spatial distribution of the direct and reverberant sound energy plays an important role in this regard. Furthermore, the DRR determined from omni-directionally measured RIRs will differ from the energy ratio arriving at the left and the right ear. An in-depth analysis of binaural DRR values is of interest.

Early-To-Late-Energy-Ratio (ELR) The DRR is a special case of measures for the *Early-To-Late-Energy-Ratio (ELR)*. Other common parameters additionally consider room reflections arriving within a certain time range after the direct sound as early energy. This is interesting due to the temporal integration of direct sound and early reflections. An example is the *Definition* D_50 (originally German 'Deutlichkeit') [10]. It was developed by Meyer and Thiele [299] and can be calculated from the impulse response as follows:

$$D_{50} = \frac{\int_{0}^{50ms} h^2(t)dt}{\int_{0}^{\infty} h^2(t)dt} * 100\% \tag{B.4}$$

Both integrals include the direct sound which arrives at the listener at $t = 0$. D_{50} has a high correlation with 'syllable intelligibility'.

The *clarity index* C_{80} (originally German 'Klarheitsmaß') was developed by Reichardt et al. [300] in order to assess the transparency of music in a concert hall.

$$C_{80} = 10log_{10} \frac{\int_{0}^{80ms} h^2(t)dt}{\int_{80ms}^{\infty} h^2(t)dt} dB \tag{B.5}$$

The longer time range after the direct sound was chosen because 'with music a reflection from a room wall is less perceptible than it is with a speech signal' [10, p. 169]. Bradley [301] showed that C_{80} is also a good predictor of the speech intelligibility. However, C_{50} is preferable with speech,

because a time range of 50 ms instead of 80 ms takes into account that with speech room perception is different than with music.

'The assumption of a sharp delay limit separating useful from non-useful reflections is certainly a crude approximation to the way in which repetitions of sound signals are processed by our hearing' [10, p. 169]. Additionally, a slightly increased delay of a strong reflection could lead to a strong change in the defined parameters. For this reason, according to [10] a linear transition between early and late was proposed by various authors, for example over a time range from 35 ms to 100 ms after the direct sound.

A parameter that does not require a division of the impulse response into early and late part is the *Center Time* t_s or center-of-gravity time t_g [6]. It was introduced by Kürer [302] and is defined as:

$$t_s = \frac{\int_0^\infty h^2(t)\, t\, dt}{\int_0^\infty h^2(t) dt} \tag{B.6}$$

It is the center of gravity of the impulse response's energy over time. Low values of t_s correspond to high transparency or speech intelligibility.

The standard DIN EN ISO 3382-1 [51] states a JND of 1 dB for C_{80}, 0.05 for D_{50} and a JND of 10 ms for t_s. Martellotta [94] argues that JNDs for Clarity have mostly been investigated with regard to rooms for speech, which usually exhibit reverberation times RT below 2 s. Therefore, he investigated JNDs for C_{80} and t_s for churches with RT between 2 s and 6 s with two different pieces of music. For C_{80}, the observed JNDs were basically independent of RT and the type of music with 1.5 ± 0.1 dB. Conversely, the JNDs for t_s varied with RT. In addition, Martellotta points out, that only 13 of the 40 participants were consistent in their evaluations. Thus, with a panel of expert listeners, smaller JND values might be found.

Investigations regarding the JNDs of C_{80}, C_{50} and t_s can be found in [94], [303] and [304].

Center Time (t_s) A parameter that does not require a division of the impulse response into early and late part is the *Centre Time* t_s. It is defined as:

$$t_s = \frac{\int_0^\infty h^2(t)\, t\, dt}{\int_0^\infty h^2(t) dt} \tag{B.7}$$

Low values of t_s correspond to high transparency or speech intelligibility.

Lateral Energy Fraction (LF) In 1967, Marshall [305] first noted the importance of lateral early reflection for the "spatial responsiveness" that is in favor of a good auditory impression of a concert hall. Barron and Marshall [124] continued to study the perceptual effect of early lateral reflections and derived *early lateral energy fraction* as "a physical measure for spatial impression" [124, p. 229] of concert halls.

An estimation according to [51] requires a calibrated measurement with an omni-directional and a figure-of-8 microphone. The sensitivity of both microphones needs to be matched. The null of the figure-of-8 microphone has to be pointed towards the sound source. The early lateral energy fraction can be determined from the measured impulse responses as follows:

$$LF_{80} = \frac{\int_{0.005}^{0.080} p_L^2(t) dt}{\int_0^{0.080} p^2(t) dt} \tag{B.8}$$

where $p_L(t)$ is an impulse response measured with a figure-of-8 microphone, $p(t)$ is measured with an omni-directional microphone.

The Early Lateral Energy Fraction is associated with the apparent source width ASW, Late Lateral Energy Fraction relates to Listener Envelopment. According to the standard [51] typical values of the Early Lateral Energy Fraction range within 0.05 and 0.35 and the just-noticeable difference is 0.05. It should be determined separately from different frequency bands. In order determine the final value according to the standard, an average is calculated over the range from 125 to 1000 Hz.

Inter-aural cross correlation coefficients (IACC) The Interaural Cross-Correlation IACC provides information about the similarity between the signals of the left and the right ear. According to the standard [51] the Inter-aural cross correlation function (IACF) is calculated from the two channels s_{left} and s_{right} of the binaural room impulse response with the following equation.

$$IACF(\tau) = \frac{\int_{t_1}^{t_2} s_{left}(t) s_{right}(t+\tau) dt}{\sqrt{\int_{t_1}^{t_2} s_{left}^2(t) dt \int_{t_1}^{t_2} s_{right}^2(t) dt}} \tag{B.9}$$

The IACC is the maximum of the magnitude within the given time range.

$$IACC_{t_1,t_2} = max|IACF_{t_1,t_2}(\tau)| \qquad for -1\,ms < \tau < +1\,ms \tag{B.10}$$

According to the standard [51] the JND of the IACC is 0.075. However, it is known, that the JNDs of the IACC depend strongly on the initial IACC [306]. Durlach [307] showed that for narrow-band noise with an initially high IACC, the sensitivity to changes to very high, resulting in low JNDs. If the initial IACC is low, listeners exhibit a low sensitivity to a change in the IACC [308].

'For narrow-band noise the sensitivity to changes in correlation is very high provided that the initial correlation is high [307]. When the correlation is low, however, subjects are rather insensitive to changes in correlation [308]. This leads to smaller JNDs for correlation differences towards a higher IACC and larger JNDs towards uncorrelated signals [306]' [93, p. 1].

Moreover, the results of Klockgether and van de Par [93] *'indicate that JNDs in the basic spatial cues (ITDs, ILDs and IACC) depend highly on the room acoustics and also on the nature of the used sound source'*.

The standard recommends to determine the IACC values in octave bands between 125 Hz and 4000 Hz. However, an analysis in 1/3 octave-bands seems to be in better correlation with the perception [309]. The standard does not provide a recommendation about how to chose t_1 and t_2. The IACC relates to two different types of spatial impression. Apparent Source Width (ASW) is associated with the $IACC_E$ of the early reflections, while the $IACC_L$ of the late reverberation seems to be a good indicator for Listener Envelopment (LEV).

Calculate room acoustic parameters from BRIRs instead of RIRs - what is the effect

The standard proposes the calculation of RT, C_{80} etc. from omni-directionally measured impulse responses. Would we get the same results by a calculation based on BRIRs?

Hak et al. [310] investigated the this question with impulse responses measured in a concert hall. They documented the following observations:

Starting from a single source position and a large number of measurements at two receiver positions with an omni-directional microphone and a HATS in a concert hall, the following can be concluded:

- Measurements with a HATS show a clear directivity of the hall for the Strength G, the Early Decay Time EDT, the Reverberation Time T30 and the Clarity C80, for all measured frequency bands.

- In all cases, the Reverberation Time T 30 measured with a HATS deviates less than 3.5 % (< 0.35 JND) from the value measured with an omni-directional microphone. The other parameters in many cases result in a deviation exceeding the JND.
- Listening under any direction to the stage (with a single source position) can result in differences in the Strength G and the Clarity C80 of more than 2 dB within the stage viewing angle.
- For the Strength G and the Clarity C 80 differences between the ears can reach up to 8 dB, depending on listening direction and frequency.
- For most omni-directional measured acoustical parameters, it seems that it is not possible to use the HATS instead of the omni-directional microphone except for the T30.