The Availability of a Hidden Real Reference Affects the Plausibility of Position-Dynamic Auditory AR

Annika Neidhardt* and Anna Maria Zerlik

Institute of Media Technology, Technische Universität Ilmenau, Ilmenau, Germany

This study examines the plausibility of Auditory Augmented Reality (AAR) realized with position-dynamic binaural synthesis over headphones. An established method to evaluate the plausibility of AAR asks participants to decide whether they are listening to the virtual or real version of the sound object. To date, this method has only been used to evaluate AAR systems for seated listeners. The AAR realization examined in this study instead allows listeners to turn to arbitrary directions and walk towards, past, and away from a real loudspeaker that reproduced sound only virtually. The experiment was conducted in two parts. In the first part, the subjects were asked whether they are listening to the real or the virtual version, not knowing that it was always the virtual version. In the second part, the real versions of the scenes where the loudspeaker actually reproduced sound were added. Two different source positions, three different test stimuli, and two different sound levels were considered. Seventeen volunteers, including five experts, participated. In the first part, none of the participants noticed that the virtual reproduction was active throughout the different test scenes. The inexperienced listeners tended to accept the virtual reproduction as real, while experts distributed their answers approximately equally. In the second part, experts could identify the virtual version quite reliably. For inexperienced listeners, the individual results varied enormously. Since the presence of the headphones influences the perception of the real sound field, this shadowing effect had to be considered in the creation of the virtual sound source as well. This requirement still limits test methods considering the real version in its ecological validity. Although the results indicate that the availability of a hidden real reference leads to a more critical evaluation, it is crucial to be aware that the presence of the headphones slightly distorts the reference. This issue seems more vital to the plausibility estimates achieved with this evaluation method than the increased freedom in motion.

Keywords: auditory augmented reality, binaural synthesis, six degrees of freedom, perceptual evaluation, plausibility, authenticity, internal reference

1 INTRODUCTION

Augmented Reality (AR) aims at adding virtual elements to the real environment (Azuma, 1997; Sicaru et al., 2018). Auditory Augmented Reality (AAR) describes the enrichment of a listener's actual environment with virtual sound sources or other virtual acoustic elements like reflectors or obstacles causing acoustic shadows. A common approach to realize AAR is to use dynamic binaural synthesis over headphones or hearables (Jot and Lee, 2016; Russell et al., 2016; Garí et al., 2019;

Nagele et al., 2021). In such reproduction, the position and orientation of the listener's head are tracked, and the headphone signals are adjusted by convolving the dry mono source signal with the corresponding binaural room impulse responses (BRIR) without a noticeable delay (Lindau, 2009; Brandenburg et al., 2020). A BRIR filter characterizes the transfer path of the sound from the sound source through the room to both ears of the listener or as a substitute head (and torso) simulator with microphones in the ears. BRIRs vary with the position and orientation of the source and receiver in the room. For consideration of source or listener motion, BRIR filters have to be updated regularly and rapidly (Neidhardt et al., 2018; Wefers and Vorländer, 2018). With the goal to realize such an AAR reproduction with low-cost devices (e.g., Heller et al. (2016)), 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 important constructs.

According to Blauert (1997), 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 (2009, 2018) has proposed the plausibility illusion as one of the key 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 (2011) has adopted the construct to evaluate acoustic reproductions. According to this proposal, plausibility describes the agreement of the perceived acoustic scene with the listener's internal reference. This internal reference is basically the expectation that results from a person's individual listening experience.

Latoschick and Wienrich (2021) have argued 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" [p. 5]. Rather than assuming an illusion of plausibility, like Slater (2009) and Skarbez et al. (2017), they have defined 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" [p. 5]. In addition, Latoschick and Wienrich (2021) have proposed 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 considered plausible if it fulfills the listener's expectations. Slater (2009) and Skarbez et al. (2017) have stated that a virtual element can be plausible even if the user knows it is not real. However, if a virtual replicate of a real sound object is in satisfactory agreement

with the individual expectations of the listener, this listener will not be able to tell for sure 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 also 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 (Moore et al., 2010; Maseiro, 2012; Oberem et al., 2016). Brinkmann et al. (2017) have 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 based on individual BRIR measurements. For these measurements, extra-aural headphones (Erbes et al., 2012) were placed over the ears of the listener to consider their influence on listening to the real scene. An experiment with an individual two-alternative forced choice (2AFC) test paradigm was conducted to test for small 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 (2012) have 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, e.g., by Hartmann and Wittenberg (1996) to evaluate externalization and convincingness, by Langendijk and Bronkhorst (2000) to investigate the *fidelity* of virtual sound sources, and in an earlier study by Lindau et al. (2007). However, Lindau and Weinzierl (2012) have taken this approach to a new level of depth and linked it to plausibility, as proposed by Kuhn-Rahloff (2011).

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 also added to the virtual version to avoid this occlusion or shadowing effect causes 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. In the investigation of a 6DOF-system, this 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 (2012) is valid and interesting for evaluating the reproduction system itself. However, reproduction systems need to be tested for plausibility, as well as fictional scenes or other contents for which there are no real counterparts. 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. (2020); Remaggi et al. (2019).

In the field of VR, scientists have started to distinguish between internal and external plausibility. Hofer et al. (2020) have provided 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" [p. 2]. An example of violated internal plausibility, as defined by Hofer et al., would be to have a vegetarian that eats meat in the scene because the new information—the character eats meat—contradicts the already presented information—the character is a vegetarian. External plausibility in this context "refers to how consistent the virtual environment is to user's 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, that is external plausibility as described by Hofer et al. 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 contents.

In three previous studies (Neidhardt et al., 2018; Kamandi, 2019; Neidhardt and Knoop, 2017), we have evaluated the plausibility of an interactive approaching motion towards a virtual sound source without considering a real scene. The participants were asked to rate plausibility directly with the four answering options "clearly plausible," "rather plausible," "rather not plausible," and "clearly not plausible." In all these studies, the position-dynamic binaural synthesis was realized with the same reproduction setup to create the spatial auditory illusion. Each study included at least one test case with a BRIR dataset fully measured in the corresponding room. In all studies, this fully measured scene was rated as plausible by all participants. Alongside plausibility, Neidhardt et al. (2018) and Kamandi (2019) have asked for continuity, externalization, sound source stability, and the impression of walking towards a sound source. In both experiments, the plausibility ratings varied substantially

according to the degree of simplification of the selected test scenes. The results for plausibility show quite a strong correlation with all of the four other attributes. In contrast, for example, continuity and externalization, or externalization and sound source stability, exhibit very low correlation. This suggests that asking directly for plausibility provides a suitable evaluation of the overall impression of the spatial auditory illusion. Our previous studies provide meaningful evaluations of the plausibility of dynamic binaural walk-through scenarios, although no real counterpart was included in the test. However, we want to know how our system performs in an experiment taking the real version into account. Generally, it is of interest how the results of an evaluation in the two different paradigms compare. Would they lead to the same conclusion?

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, e.g., indicating the agreement with the pure internal reference or the tuned internal reference resulting from listening to the real version of the scenario. **Table 1** summarizes a selection of previous studies on the authenticity and the two proposed categories of 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 quite new challenge concerning the evaluation of plausibility.

A potential tuning of the internal reference may occur in an indirect comparison with the real counterpart. Especially for AAR, the actual environment and its components are likely to 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. (2020) have 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 suggested 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. (2018) addresses this topic similarly.

It is likely that the number of sources or the scene complexity, as well as the type and the relative positions of the available real sound sources, influences the internal reference. 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.

With this new study, we want to evaluate our position-dynamic AAR system with the approach proposed by Lindau and Weinzierl. To our knowledge, this is the first time this

TABLE 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.

Binaural synthesis	Plausibility I pure internal reference	Plausibility II "tuned" internal reference	Authenticity external reference
Static reproduction	(✓)	Hartmann and 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 and Weinzierl (2012) Pike et al. (2014)	Brinkmann et al. (2017)
Rotation and translation	Neidhardt and Knoop (2017) Neidhardt et al. (2018)		
	This study, part I	This study, part II	

approach is applied to a system that provides interactive walking. Furthermore, it is of interest to estimate the relevance of including the real version in evaluating plausibility. Therefore, we created an experiment to assess the plausibility of the auditory illusions created with our AAR system with and without real versions of the scenes among the test items. The following section presents the technical realization of the evaluated AAR system, the test scenario chosen for the experiment, and the test design.

2 MATERIALS AND METHODS

The test scenario was realized in a seminar room of the university in Ilmenau. The participants had to wear headphones. The two loudspeakers standing in the room could reproduce sound either 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.

2.1 Choice of Headphones

Satongar et al. (2015) have shown that the passive influence of headphones can cause spectral distortions, affect the effective interaural time difference, and reduce localization accuracy. Brinkmann et al. (2017) have used the extra-aural headphones BK211 presented by Erbes et al. (2012) for their experiment on authenticity. These headphones may be the best choice for a mixed-reality scenario with respect to 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 effort to attach them stably to the listener. It may be assumed that wearing these headphones does not allow for a perfectly natural motion. Especially during walking, people may move more carefully to avoid changing the headphone position on the head. For this reason, we decided not to use the extra-aural headphones in this experiment.

Lindau and Weinzierl (2012) and Pike et al. (2014) have 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. **Figure 1** shows the setup. In the aftermath of this study, we analyzed these effects for different headphones, including all the mentioned ones (Schneiderwind et al., 2021). Our discussion considers these results.

2.2 Measurement of Binaural Room Impulse Responses

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 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 and one 1.25 m right of the line as illustrated in **Figure 2**. 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 same 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 (2009). The captured BRIRs and the created headphone compensation filter are provided as an open-access dataset by Neidhardt (2019).
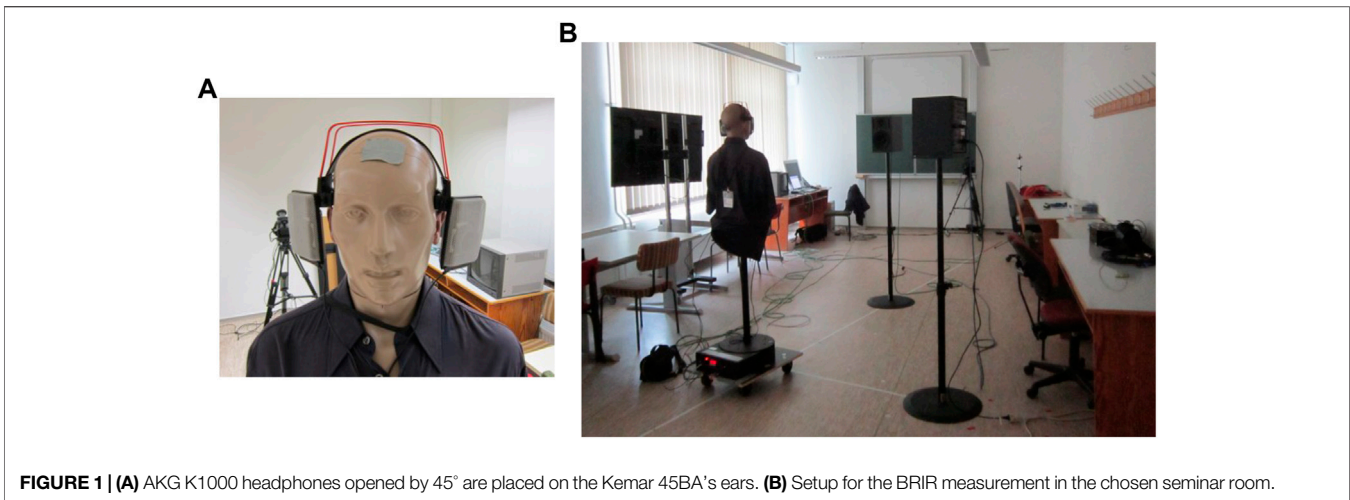
FIGURE 1 | **(A)** AKG K1000 headphones opened by 45° are placed on the Kemar 45BA's ears. **(B)** Setup for the BRIR measurement in the chosen seminar room.
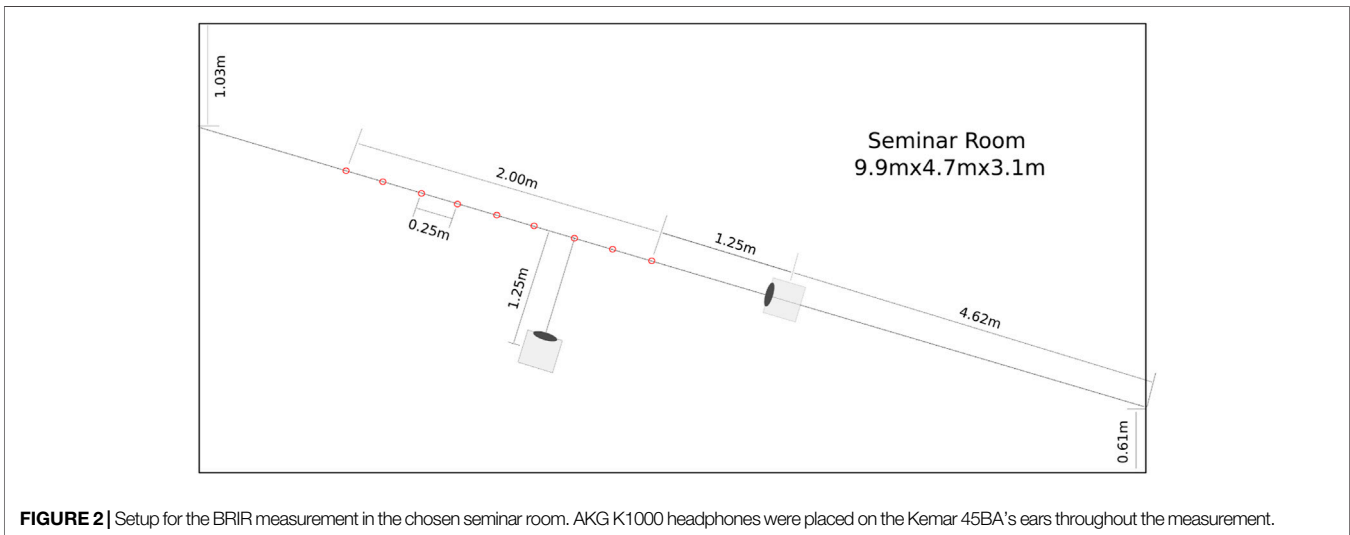
FIGURE 2 | Setup for the BRIR measurement in the chosen seminar room. AKG K1000 headphones were placed on the Kemar 45BA's ears throughout the measurement.

2.3 Position-Dynamic Reproduction Setup for Auditory AR

After the measurement, the two loudspeakers were kept in exactly the same positions of the same room. An HTC Vive tracker was attached to the headphones to track the position and orientation of the listener's head, as shown in **Figure 3**. 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. (2017) 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 65,536 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 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.

2.4 Individualization of Binaural Audio

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

FIGURE 3 | Test person wearing AKG K1000 headphones with a Vive tracker attached to them.

TABLE 2 | Two different source positions, three types of signals, and two different sound levels were taken into account.

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

time difference (ITD) combined with an adequate BRIR adjustment. Brinkmann et al. (2017) have measured individual BRIRs for each participant before evaluating the authenticity of the binaural reproduction. Lindau and Weinzierl (2012) have 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., 2007). Pike et al. (2014) have 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 (2012). 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 described as increased localization blur or increased apparent source width. In their experiment, a sensory distance between real sound field and auralization was found.

In a test paradigm without considering a real scene, dynamic binaural synthesis with non-individual BRIRs was repeatedly perceived as plausible (Neidhardt and Knoop, 2017; Neidhardt et al., 2018; Kamandi, 2019).

2.5 Participants

Seventeen people aged between 18 and 33 years volunteered for participation in the experiment. The average age was 25 ± 2.57 years. Five of the subjects were experienced listeners in the field of BRIR-based binaural synthesis, and the others were mostly inexperienced. Experienced listeners were expected to be more critical about 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 students or Ph.D. candidates at the university in Ilmenau and 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 full experiment and all their results were included in the statistical analysis.

2.6 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. **Table 2** 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.

2.7 Pre-Test with Few Experts

In the preparation of 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 **section 3**. In the course of 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.

2.8 Listening Experiment With the Test Panel

Before participating in the study, informed consent was obtained from all individual participants involved in the study. In the

experiment, the participant had to wear the AKG K1000 headphones with the Vive tracker attached to them. 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 randomly either the real loudspeaker or its binaural simulation would be presented, and 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 the plausibility with respect to the 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 min 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 min per participant.

The participants were asked to evaluate 36 test scenes wherein the number 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. The test procedure was designed in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

2.9 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 take the decision may affect the rate of correct answers. Lindau et al. have 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 had a duration of only 6 seconds, which was possible because interactive self-motion was limited to $\pm80°$ 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 s per test scene. Between the scenes, there was a break of 20–25 s for the test conductor to take notes and start the new scene. In total, the experiment with introduction and interview at the end took between 50 and

TABLE 3 | Possible outcomes in the Signal Detection paradigm.

Response	Virtual (signal)	Real (noise)
"Virtual"	Hit	False Alarm (FA)
"Real"	Miss	Correct Rejection (CR)

70 min. Due to the breaks, the active exploration of the scene, and the reportedly interesting task, listener fatigue was kept at an acceptable level.

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

2.10 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.
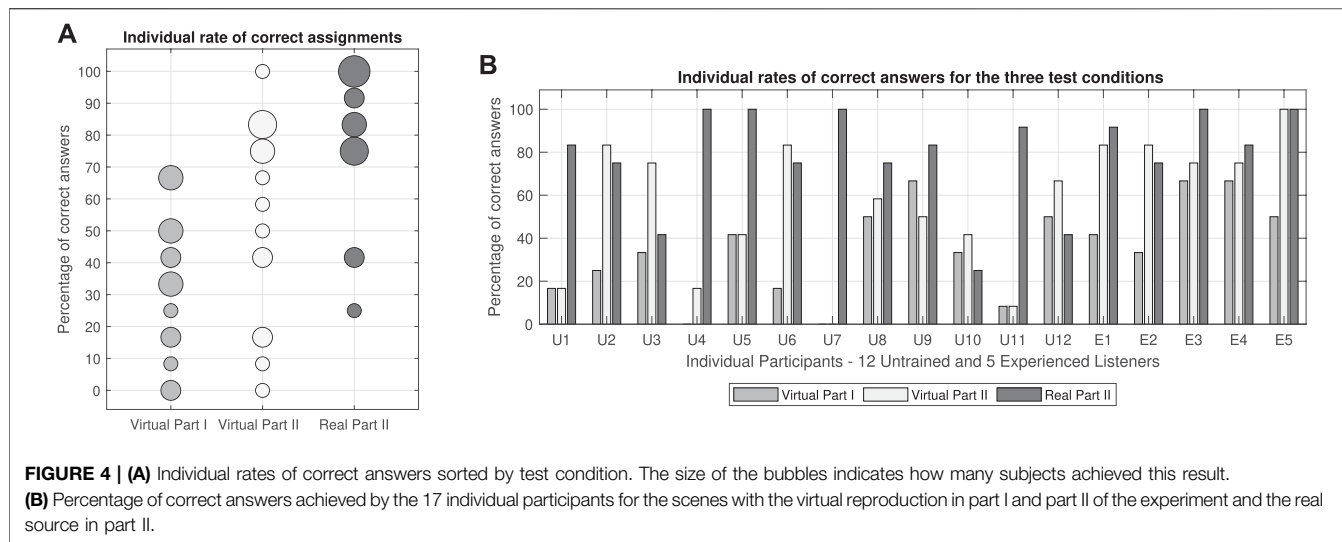
2.10.1 Estimating the Discriminability Based on Signal Detection Theory

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 with Herzog et al. (2019), the four possible outcomes in this classic SDT experiment are called Hit, Miss, False Alarm, and Correct Rejection. **Table 3** provides an overview.

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 for 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 version. If more subjects and more stimuli are taken into account, the theory demands determining the individual sensitivity d' for each combination of subject and stimulus separately and then calculating 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 in Macmillan and Creelman (2004) [$p.$ 331 ff].

Several previous plausibility studies have used the SDT for their analysis. For example, Lindau and Weinzierl (2012) have calculated the individual sensitivity per person, averaging over different signals and source positions, then calculated the mean

FIGURE 4 | (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 and the real source in part II.

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 considered in the SDT offers Noise and Signal Stimuli (in our experiment, real and virtual) within one trial, in either randomized temporal or spatial order. Therefore, a 2AFC paradigm allows for direct comparison between both stimuli. Furthermore, the answer in each trial is correct or wrong for both stimuli at the same time. In contrast, in a Yes/No paradigm, distinguishing Hits and Correct Rejections can provide additional or more accurate information since they are not necessarily equal. **Figure 4** visualizes the individual 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., Wickens (2001), p_{Hit} and p_{FA} can be calculated as follows:

$$p_{Hit} = \frac{\text{Number of Hits}}{\text{Total Number of Signal Presentations}}$$

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

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

$$d' = z(p_{Hit}) - z(p_{FA}) \tag{1}$$

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 (1995) was applied. This correction is integrated into the *dprime*-function in R, which we used for this analysis. Since the sample size per person is 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.

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

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 (2012) have determined such a minimum effect hypothesis under the assumption of non-biased participants and only considering the percentage of correct answers. For a group of subjects with considerable differences in individual bias, the determination becomes more challenging. Therefore, we additionally consider another interpretation of the data.

2.10.2 Analysis Based on the Paired *t*-Test

It is interesting to analyze the rate of acceptance as real. For the real source, this number is equivalent to the number of the correct answers. For virtual reproduction, it is the number of wrong answers. The auditory illusion can be considered plausible if the rate of acceptance 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 testing 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.

3 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 (2012) by including real versions of the simulated sound fields. Does the availability of the real sound field affect the plausibility?

3.1 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 **Table 2**. The experts described freely which differences they perceived. It was interesting to notice that after a short episode of exploration, the experts moved to the closest position possible to the front of the active sound source. Once they arrived there, they focused on rotating their heads or turning themselves at that position. Sometimes, they reported a slight instability of the perceived location of the sound source during head rotation. Furthermore, when turning the back towards the sound source, differences between real and virtual reproduction were audible. The deviations were described as a change in distance perception, externalization, and relative sound level. For the binaural reproduction, the source was described to be in the head or sticking to the back of the head. However, with the real reproduction, the source in the back did not appear fully natural as well. 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.

3.2 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 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." For each of these categories, each of the 17 participants rated 12 test scenes and achieved an individual number x_i of correct answers between 0 and 12.

Figure 4B illustrates the individual rates of correct answers each of the participants 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. **Figure 4A** 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, $t(16) = 3.50$, $p = 0.003$ ($M = 6.76$, $SD = 3.68$), than for the same test scenes in part I ($M = 4.24$, $SD = 2.63$).

3.3 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, at least for some of the listeners, the internal reference is not perfectly reliable. 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 which 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%).

TABLE 4 | 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.

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)

Inexperienced listeners were correct in 74.3% of the cases. **Figure 4** visualizes the individual results. Three inexperienced listeners rated the real loudspeaker reproduction as real only in three or five of the 12 test cases. Especially, the person with the three correct identifications tended to assign virtual and real scenes vice versa.

3.4 Analysis of Part 2: 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 is in accordance with the method proposed by Lindau and Weinzierl (2012). They have 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.

3.4.1 Estimating the Discriminability Based on Signal Detection Theory

The sensitivity index d' can be calculated with **Eq. (1)**. Due to the small sample size per person small, in addition to the common mean sensitivity, we determined the pooled sensitivity to compare both. The first column in **Table 4** 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. (2)**. Due to the small sample size, also c was 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."

3.4.2 Analysis Based on the Paired *t*-Test

Figure 5 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, $t(16) = 4.18$, $p < 0.001$. This means, in part II, the acceptance of the virtual reproduction ($M = 5.23$,

$SD = 3.68$) was significantly lower than that of the real reproduction ($M = 9.47$, $SD = 2.74$).

In addition to the results of the whole group of participants, **Figure 5** shows the separate results for experienced and inexperienced listeners. For both groups, the paired *t*-test separately still 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 s 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.

3.5 Analysis of Part I: Plausibility With Regard to the Pure Internal Reference

In **Figure 4B**, the first and second row of bubbles show the individual percentage of correct identifications of the virtual sound source in the first and the second part of the experiment. In part I, 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. In this case, it was identified correctly in 115 of the 204 scene assignments (56.4%). The statistical analysis is again based on the two approaches, Signal Detection Theory and the paired *t*-test.
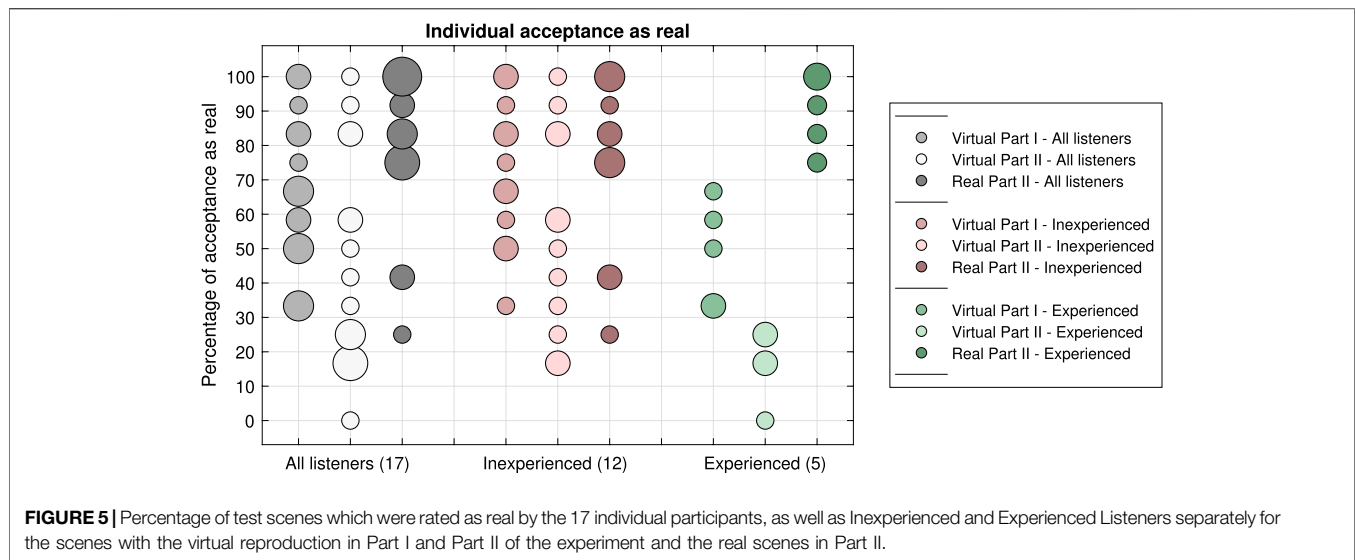
3.5.1 Analysis Based on Signal Detection Theory

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. **Table 4** provides an overview of the estimated sensitivities.

Again, mean and pooled sensitivity are very similar. As expected, the sensitivity estimated for part I is considerably lower than that 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.

3.5.2 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

FIGURE 5 | 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.

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 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 of interest 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 **Figure 5** 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 that 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.

3.6 Cues Used for Detection of the Virtual Reproduction

Figure 6 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 towards 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.

3.7 Source Position, Type of Signal, and Sound Level

Figure 7 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 number of test cases per person was twelve, six real and six virtual. According to the paired t-test, the percentage of correct answers for the source in the front ($M = 7.94$, $SD = 2.38$) does not differ significantly, $t(16) = 1.0$, $p > 0.3$, 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, $t(16) = 0.57$, $p > 0.3$, from that for the −6 dB ($M = 8.0$, $SD = 2.29$).
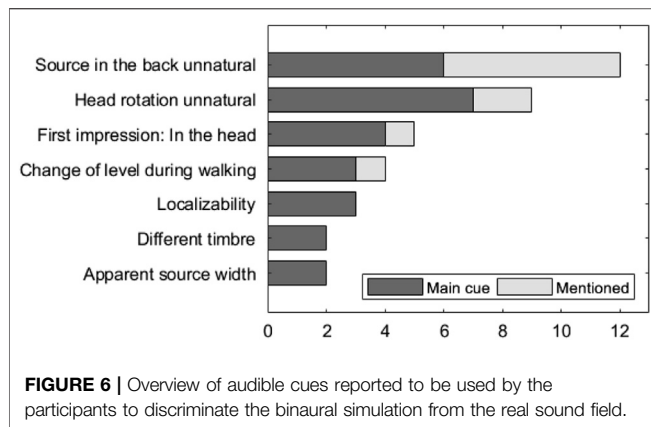
FIGURE 6 | Overview of audible cues reported to be used by the participants to discriminate the binaural simulation from the real sound field.

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 > 0.6, 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.

For the main 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 quite small for all of them together. However, a pooled inspection is possible. **Figure 7** 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 percentage correct for each condition, the confidence intervals proposed by Clopper and Pearson (1934) are shown. The virtual sources were accepted as real significantly less often than the real source for each of the conditions. 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 source position nor the level or type of signal had a significant impact on 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 differences between the test conditions did not exhibit different quality in terms of plausibility as the agreement with the tuned internal reference.

4 DISCUSSION

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.
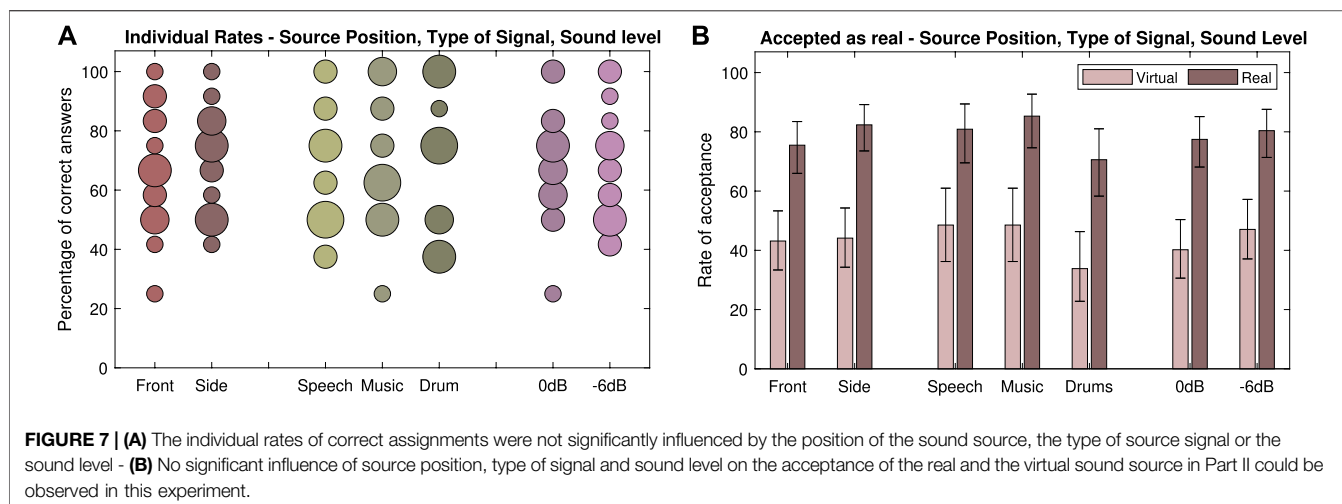
4.1 Plausibility of Position-Dynamic AAR Realization

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 main cues to identify the auralization was the audible difference in case the listener turned his back towards the source. Distance perception, externalization, and timbre were affected. All the previous studies did not document such an effect. Brinkmann et al. (2017) have tested the authenticity for source directions of 0° and 90°, allowing a head rotation of ±34°. The study was conducted with the extra-aural headphones. Lindau and Weinzierl have worked with STAX headphones and allowed a head rotation of ±80°. Pike et al. (2014) have 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 the previous studies, AKG K1000 headphones were not used. Satongar et al. (2015) have shown that the passive influence of headphones can 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. (2019) and Schneiderwind et al. (2021) 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 and Weinzierl Lindau et al. (2007) testing an early-stage system, as well as by Pike et al. (2014). This audible effect could be due to non-individualized ITDs or a non-optimal delay in the motion-related updating of the BRIR filters. These aspects have to 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 starting to move. They assigned this experience to the binaural simulation. However, in-the-head localization can occur in real sound fields as well (Plenge, 1972). It is questionable whether this is a reliable cue for the identification of virtual sound

FIGURE 7 | (A) The individual rates of correct assignments were not significantly influenced 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.

sources. Still, it may occur more often or more pronounced in a binaural reproduction.

Four participants stated that for them, the change of level during walking was a helpful cue. 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 Neidhardt et al. (2018); Kamandi (2019). Therefore, several untrained listeners were surprised about the progress of the sound level in the measured scenario and rated 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 that 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 to be 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 of 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.

The majority of the cues reported as helpful for identifying the virtual version were not related to translation. Four of the untrained subjects mentioned that the sound level would exhibit unexpected progress during walking. Similar statements were given in a previous experiment by (Kamandi, 2019) for the measured scene by participants who rated another artificial scene with a considerably greater change of the level as plausible. This judgment may be the result of an inaccurate or wrong internal reference. 13 of the 17 subjects in the present experiment did not mention any translation-related cues at all. 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, the observation regarding 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.

4.2 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 the 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, it is possible that the identification of helpful cues revealing the virtual scene is 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 in the course of 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 at 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 actually tend to answer "virtual." However, this is only an issue if a subject cannot confidently identify virtual reproduction. In contrast, at least several of the inexperienced listeners answered with real very often. Apparently, 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. The percentage of correct answers and thus the rate of acceptance 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 the 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 the main findings of this experiment are not affected by this decision.

The results of this experiment suggest that including the real version of the scenes affects the listener's capability of identifying 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 the internal reference, which is fundamental for evaluating plausibility.

The fact that including the real version affects the estimated plausibility and reduces the acceptance of the 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. (2008). The observations indicate that in the future, discriminating between different kinds of plausibility may be of interest. 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 kind of reference for a virtual acoustic element. For evaluating its quality, it is important 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 therefore even more sensitive.

4.3 How Should the Plausibility of Auditory AR Be Evaluated in the Future?

This study considers an AAR scene, which contains one primary sound source besides the common 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 towards 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. There is usually no option in such scenarios to listen to exactly 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. (2020) have already shown 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 contents in terms of 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 a consideration of 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 main cue for identifying the virtual reproduction among the real scenes was probably caused by the shadowing effect of the headphones. This raises the question, 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 some sort of 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 as in the real world without the slight influences of any headphones.

4.4 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 of them reproduced sound either 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 (2012), was applied to binaural walk-through scenarios with a true 360° experience for the first time. Including real sound scenes as test items is accompanied by some challenges and limitations. On the one hand, the method can only consider the real scene as it is perceived through the used headphones or hearables. On the other hand, these effects have to be considered in the creation of the 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 the creation of 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 not available. Even the experienced listeners could not confidently identify the presentation as a simulation in this case. 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 the source signal, and its sound level did not significantly influence the observations. Two primary cues revealed the virtual reproduction. In the listener's back, the sound source exhibited an unnatural appearance, which was caused by the presence of the headphones. In addition, the participants reported slight instabilities of the sound source during head rotation, which were probably caused by

the lack of individualization and maybe a non-optimal system latency.

4.5 Conclusion

The results of the presented study indicate that the system under test is capable of inducing a plausible illusion for inexperienced listeners. However, the system fails to deliver a plausible illusion for experienced listeners in general and for 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 demands considering the occlusion effect of the headphones 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.

DATA AVAILABILITY STATEMENT

The dataset of measured BRIRs used in this study can be found at Zenodo *https://zenodo.org/record/3457782*. The pyBinSim software Neidhardt et al. (2017) for dynamic binaural auralization is available at https://github.com/pyBinSim or https://pypi.org/project/pybinsim/.

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

As AN's research focuses on the perceptual issues of listener translation motion and an efficient realization of dynamic

binaural synthesis for an exploration in six degrees of freedom, she initiated this study and formulated the main research question. She was the supervisor of the master thesis in which this study was conducted and she also wrote the text in this article. AMZ realized the measurements and conducted the experiment as main task in her master thesis. In addition, she contributed considerably to the review of previous experiments and in taking decisions for many details of the test design. The final version of her thesis inspired several of the sections and the discussion in this article.

REFERENCES

Azuma, R. T. (1997). A Survey of Augmented Reality. *Presence: Teleoperators Virtual Environ.* 6 (4), 355–385. doi:10.1162/pres.1997.6.4.355

Blauert, J. (1997). "Spatial Hearing," in *The Psychophysics of Human Sound Localization* (Cambridge, MA, USA: MIT Press). revised edition.

Brandenburg, K., Klein, F., Neidhardt, A., Sloma, U., and Werner, S. (2020). "Creating Auditory Illusions with Binaural Technology," in *The Technology of Binaural Understanding*. Editors J. Braasch and J. Blauert (Cham, Switzerland: Springer Nature Switzerland AG), 623–663. doi:10.1007/978-3-030-00386-9_21

Brinkmann, F., Lindau, A., and Weinzierl, S. (2017). On the Authenticity of Individual Dynamic Binaural Synthesis. *J. Acoust. Soc. Am.* 142, 1784–1795. doi:10.1121/1.5005606

Clopper, C. J., and Pearson, E. S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* 26, 404–413. doi:10.2307/233198610.1093/biomet/26.4.404

Enge, K., Frank, M., and Höldrich, R. (2020). "Listening experiment on the Plausibility of Acoustic Modeling in Virtual Reality," in 46th Annual Meeting on Acoustics (DAGA), Hannover, Germany, March 16–19, 2020.

Erbes, V., Schultz, F., Lindau, A., and Weinzierl, S. (2012). "An Extraaural Headphone System for Optimized Binaural Reproduction," in In 38th Annual Meeting on Acoustics (DAGA), Darmstadt, Germany, March 19–22, 2012.

Garí, S., Brimijoin, W., Hassanger, H., and Robinson, P. (2019). "Flexible Binaural Resynthesis of Room Impulse Responses for Augmented Reality Research," in In EAA Conference on Spatial Audio Signal Processing, Paris, France, September 6–7, 2019.

Hartmann, W. M., and Wittenberg, A. (1996). On the Externalization of Sound Images. *J. Acoust. Soc. Am.* 99 (6), 3678–3688. doi:10.1121/1.414965

Hautus, M. J. (1995). Corrections for Extreme Proportions and Their Biasing Effects on Estimated Values Ofd′. *Behav. Res. Methods Instr. Comput.* 27, 46–51. doi:10.3758/bf03203619

Heller, F., Jevanesan, J., Dietrich, P., and Borchers, J. (2016). "Where Are We?," in Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services, Florence, Italy, September 6–9, 2016. doi:10.1145/2935334.2935365

Herzog, M., Francis, G., and Clarke, A. (2019). *Understanding Statistics and Experimental Design - How to Not Lie with Statistics*. Germany: Springer Open Books. Learning Materials in Biosciences.

Hofer, M., Hartmann, T., Ratan, R., Eden, A., and Hahn, L. (2020). The Role of Plausibility in the Experience of Spatial Presence in Virtual Environments. *Front. VR* 1, 1–9. doi:10.3389/frvir.2020.00002

Jot, J., and Lee, K. (2016). "Augmented Reality Headphone Environment Rendering," in 2016 Aes International Conference On Audio For Virtual And Augmented Reality, Los Angeles, CA, USA, September 30–October 1, 2016.

Kamandi, S. (2019). Perception of Simplifications of the Room Acoustics in a Dynamic Binaural Synthesis for Listener Translation. Master's thesis. Germany: Technische Universität Ilmenau.

Kuhn-Rahloff, C. (2011). Prozesse der Plausibilitätsbeurteilung am Beispiel ausgewählter elektroakustischer Wiedergabesituationen. Ph.D. thesis. TU Berlin, Germany.

Langendijk, E. H. A., and Bronkhorst, A. W. (2000). Fidelity of Three-Dimensional-Sound Reproduction Using a Virtual Auditory Display. *J. Acoust. Soc. Am.* 107 (1), 528–537. doi:10.1121/1.428321

Latoschick, M., and Wienrich, C. (2021). Coherence and Plausibility, Not Presence?! Pivotal Conditions for XR Experiences and Effects, a Novel Model. *arXiv* [Epub ahead of print].

Lindau, A. (2009). "The Perception of System Latency of Dynamic Binaural Synthesis," in NAG/DAGA International Conference on Acoustics, Rotterdam, The Netherlands, March 23–26, 2009.

Lindau, A., Hohn, T., and Weinzierl, S. (2007). "Binaural Resynthesis for Comparative Studies of Acoustical Environments," in In 122nd International AES Convention, Vienna, Austria, May 5–8, 2007.

Lindau, A., and Weinzierl, S. (2012). Assessing the Plausibility of Virtual Acoustic Environments. *Acta Acustica united with Acustica* 98, 804–810. doi:10.3389/fnins.2013.1234510.3813/aaa.918562

Macmillan, N., and Creelman, C. (2004). *Detection Theory - a User's Guide*. New Jersey: Psychology Press.

Maseiro, B. (2012). Individualized Binaural Technology. Measurement, Equalization and Perceptual Evaluation. Ph.D. thesis. Germany: RWTH Aachen.

Moore, A., Tew, A., and Nicol, R. (2010). An Initial Validation of Individualized Crosstalk Cancellation Filters for Binaural Perceptual Experiments. *J. Audio Eng. Soc.* 58, 36–45. doi:10.1121/1.5005606

Nagele, A. N., Bauer, V., Healey, P. G. T., Reiss, J. D., Cooke, H., and Cowlishaw, T. (2021). Interactive Audio Augmented Reality in Participatory Performance. *Front. Virtual Real.* doi:10.3389/frvir.2020.610320

Neidhardt, A. (2019). "BRIRs for Position-Dynamic Binaural Synthesis Measured in Two Rooms," in In 5th International Conference on Spatial Audio (ICSA), Ilmenau, Germany. Data set.

Neidhardt, A., Klein, F., Knoop, N., and Köllmer, T. (2017). "Flexible python Tool for Dynamic Binaural Synthesis Applications," in In 142nd International AES Convention, Berlin, Germany, May 20–23, 2017.

Neidhardt, A., and Knoop, N. (2017). "Binaural Walk-Through Scenarios with Actual Self-Walking Using an HTC Vive," in In 43rd Annual Conference on Acoustics, Kiel, Germany, March 6–9, 2017.

Neidhardt, A., Tommy, A., and Pereppadan, A. (2018). "Plausibility of an Interactive Approaching Motion towards a Virtual Sound Source Based on Simplified BRIR Sets," in In 144h International AES Convention, Milan, Italy, May 23–26, 2018.

Oberem, J., Maseiro, B., and Fels, J. (2016). Experiments on Authenticity and Plausibility of Binaural Reproduction Ia Headphones Employing Different Recording Methods. *Appl. Acoust.* 114, 71–78. doi:10.1016/j.apacoust.2016.07.009

Pike, C., Melchior, F., and Tew, T. (2014). "Assessing the Plausibility of Non-individualized Dynamic Binaural Synthesis in a Small Room," in In 55th International AES Conference, Helsinki, Finland, August 27–29, 2014.

Plenge, G. (1972). On the Problem of "in Head Localization". *Acta Acustica united with Acustica* 26, 241–252.

Pörschmann, C., Arend, J., and Gillioz, R. (2019). "How Wearing Headgear Affects Measured Head-Related Transfer Functions," in In 1st EAA Spatial Audio Signal Processing Symposium, Paris, France, September 6–7, 2019.

Remaggi, L., Kim, H., Neidhardt, A., Hilton, A., and Jackson, P. (2019). "Perceived Quality and Spatial Impression of Room Reverberation in VR Reproduction from Measured Images and Acoustics, September 9–13, 2019.

Russell, S., Dublon, G., and Paradiso, J. (2016). "Hearthere - Networked Sensory Prosthetics through Auditory Augmented Reality," in 18th International

FUNDING

ACKNOWLEDGMENTS

Conference on Human-Computer Interaction with Mobile Devices and Services, Florence, Italy, September 6–9, 2016. doi:10.1145/2875194.2875247

Satongar, D., Pike, C., Lam, Y., and Tew, A. (2015). The Influence of Headphones on the Localization of External Loudspeaker. *J. Audio Eng. Soc.* 63 (10), 799–810. doi:10.17743/jaes.2015.0072

Schärer, Z., and Lindau, A. (2009). "Evaluation of Equalization Methods for Binaural Signals," in In 126th International AES Convention, Munich, Germany, May 7–10, 2009.

Schneiderwind, C., Neidhardt, A., and Meyer, D. (2021). "Comparing the Effect of Different Open Headphone Models on the Perception of a Real Sound Source," in In 150th International AES Convention, May 25–29, 2021. Online.

Sicaru, I., Ciocianu, C., and Boiangiu, C.-A. (2018). A Survey on Augmented Reality. *J. Inf. Syst. Operations Manag.* 11 (2), 355.

Skarbez, R., Jr., Brooks, F., and Whitton, M. (2017). A Survey of Presence and Related Concepts. *ACM Comput. Surv.* 50 (6), 1–39. doi:10.1145/3134301

Slater, M. (2018). Immersion and the Illusion of Presence in Virtual Reality. *Br. J. Psychol.* 109, 431. doi:10.1111/bjop.12305

Slater, M. (2009). Place Illusion and Plausibility Can lead to Realistic Behaviour in Immersive Virtual Environments. *Phil. Trans. R. Soc. B* 364, 3549–3557. doi:10.1098/rstb.2009.0138

Stecker, G., Moore, T., Folkerts, M., Zotkin, D., and Duraiswami, R. (2018). "Towards Objective Measures of Auditory Co-immersion in Virtual and Augmented Reality," in AES International Conference on Audio for Virtual and Augmented Reality, Redmond, WA, USA, August 20–22, 2018.

Wefers, F., and Vorländer, M. (2018). Flexible Data Structures for Dynamic Virtual Auditory Scenes. *Virtual Reality* 22, 281. doi:10.1007/s10055-018-0332-9

Wickens, T. (2001). *Elementary Signal Detection Theory*. Germany, USA: Oxford University Press.

Wirler, S., Meyer-Kahlen, N., and Schlecht, S. (2020). "Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes," in AES International Conference on Audio for Virtual and Augmented Reality, August 17–19, 2020.

Zielinski, S., Rumsey, F., and Bech, S. (2008). On Some Biases Encountered in Modern Audio Quality Listening Tests - a Review. *J. Audio Eng. Soc.* 56 (6), 427–451.