OPEN ACCESS Freely available online

T. Lübeck, T. Weber, and C. Pörschmann,
"A Perceptual Study on the Plausibility of Binaural
Rendering in Anechoic Environments,"
J. Audio Eng. Soc., vol. 72, no. 12, pp. 846–859 (2024 Dec.).
https://doi.org/10.17743/jaes.2022.0180.

A Perceptual Study on the Plausibility of Binaural Rendering in Anechoic Environments

TIM LÜBECK,[*,†] TOBIAS WEBER,[†] AND CHRISTOPH PÖRSCHMANN

(tim.luebeck@th-koeln.de) (tobias.weber1@th-koeln.de) (christoph.poerschmann@th-koeln.de)

Technische Hochschule Köln, Cologne, Germany

Plausibility has become a widely used paradigm for evaluating binaural audio for augmented reality. Although several studies have evaluated the plausibility of different binaural rendering techniques, it is not clear which parameters contribute to plausibility and to what extent. The present study attempts to fill this gap and presents the results of five listening experiments systematically investigating the impact of different parameters. A first plausibility test performed in anechoic conditions assessed the plausibility of nonindividual head-related transfer-function rendering. In four subsequent experiments, the influence of spherical harmonics truncation order, visual cues, small translational head movements, and different head-related transfer-function sets is assessed. Overall, the results indicate that even large manipulations of the rendering system have surprisingly little impact on the estimated d'.

0 INTRODUCTION

A widely used method for creating 3D audio for virtual reality (VR) and augmented reality (AR) applications is headphone-based dynamic binaural synthesis [1]. Classically, the headphone signals are synthesized by convolving dry audio content with a pair of binaural room impulse responses (BRIRs) picked according to the listener's actual head orientation. In the case of a virtual acoustic environment (VAE) with 6 degrees of freedom (DoF), the BRIRs are also selected according to the listener's position. With the advent of VR/AR devices in the consumer sector, the demand for efficient methods for binaural reproduction is becoming increasingly important. Measuring or simulating large datasets of BRIRs is mostly impossible. In addition, technical issues such as the appropriate grid resolution and appropriate interpolation methods in 3D space can be challenging [2, 3].

An alternative method for creating binaural VAEs is parametric spatial audio rendering. The sound field is parametrized (encoded) into a set of perceptually relevant parameters and then spatially decoded in real time. This technique offers high variability according to the desired application and is usually very computationally efficient. In the last few decades, a large number of parametric approaches have been introduced [4–8]. Mostly, the sound field is decomposed into parameters describing the directional part and parameters encoding the diffuse part. With a set of head-related transfer functions (HRTFs), the parameters can then be decoded binaurally in real time. A method for extrapolation and 6 DoF auralization is presented, for example, by [9, 10].

This work focuses on binaural audio for AR [auditory AR (AAR)], which aims to render virtual sound sources together with real room acoustics. The main goal of AAR is to create perceptually convincing acoustic spaces consistent with the real acoustic environment. For the perceptual evaluation of such applications, "plausibility" has become an established paradigm. According to Slater [11], two components contribute to the perceived realism of virtual environments: 1) "Presence," usually described as being there, which Slater called "Place Illusion" (Pi), and 2) "Plausible Illusion," which refers to the illusion "that what is apparently happening is really happening," or, in other words, "overall credibility of the scenario being depicted in comparison with expectations." He also discussed the relevance of Plausible Illusion and Pi for the perception of multimodal VR/AR. The concept of Pi matches the concept of "external plausibility" introduced by Hofer et al. [12]. A comprehensive summary of the presence and plausibility concepts in the context of virtual environments is presented in [13].

Similar to Slater's definition, Lindau and Weinzierl [14] described plausibility as the "agreement with the listener's

[*]To whom correspondence should be addressed, email: tim.luebeck@th-koeln.de.

[†]Tim Lübeck and Tobias Weber contributed equally to this work.

expectation towards a corresponding real event" and introduced a test method to assess the plausibility of VAEs. This procedure is based on a yes/no paradigm and has become an established method for investigating the plausibility of AAR systems. Listeners are presented with a series of stimuli, either from a real sound source, usually a loudspeaker, or synthesized via headphones using dynamic binaural synthesis. Each successive stimulus consists of a different audio content and a different source position so that the participants never listen to an explicit real-world reference. After each stimulus, participants indicate whether they believe the sound source to be real or simulated.

Initial results of such a test were also published by Lindau and Weinzierl [14], who accessed the plausibility of data-based binaural synthesis with BRIRs measured with an artficial head. This was followed by a number of studies evaluating the plausibility of different binaural synthesis techniques [15–18]. The classical plausibility test is highly sensitive to technical artifacts, such as tracking inaccuracies or slight headphone crackling. If a participant identifies such an artifact and can attribute it to the virtual reproduction, plausibility breaks down immediately. All the studies mentioned above used different setups for their rendering system, including different algorithms for binaural synthesis, different tracking systems, or headphones. Consequently, it remains uncertain which parameters lead to the different plausibility results.

This work contributes to a better understanding of which components of binaural synthesis are relevant for plausible rendering. The binaural rendering is collapsed to its fundamental elements, namely HRTFs, source directivity, and distance attenuation. Using these basic components of binaural rendering, a series of plausibility tests in anechoic conditions, systematically altering various parameters of the rendering system, was performed. All experiments strictly applied the procedure and analysis by Lindau and Weinzierl [14], enabling a direct comparison with other studies employing the same methodology at the conclusion of the article.

In a baseline test, HRTFs were applied measured with a Neumann KU100 mannequin wearing the same open headphones as the participants during the test. The corresponding HRTF was picked according to the actual head orientation and translational head movements of the sitting participants. During the test, the real loudspeakers were hidden behind a curtain such that participants only focused on the auditory cue.

Building on this baseline test, the system was varied along different axes: A first variation examined the influence of lowering the spatial resolution. Several spatial audio rendering methods perform encoding and decoding in the spherical harmonics (SH) domain, such as classical Ambisonics-based rendering of spherical microphone arrays [19], parametric renderings such as Higher-Order Spatial Impulse Response Rendering [20, 21] or the Spatial Decomposition Method [5], or several simulation approaches. Thereby, the SH order is associated with the spatial resolution. Higher spatial orders require more measurement or computational effort, which is why finding an optimal rendering order is of high interest for the VR/AR sector. Thresholds of indistinguishability have already been determined in several studies [22–25]. This work investigated the influence on plausibility and conducted three plausibility tests with HRTFs truncated at orders $N = [12, 7, 3]$.

Next, the focus is on the influence of visual information on perceived plausibility. Several studies discussed the interaction of visual and auditory cues and their interaction in AR/VR [26–29]. Recently Riedel et al. [30] showed the influence of visual cues on localization performance when using nonindividual HRTFs. Bailey and Fazenda [27] conducted a plausibility test with different ways of presenting a visual representation of the sound sources through a head-mounted display. They presented a virtual environment including visible loudspeakers, visual environment without visible loudspeakers, or fully black screen. They concluded that the absence of visible sound sources in the virtual environment leads to a reduced number of correctly identified presentations. To evaluate how visual information affects plausibility in a scenario with real visible sound sources, the curtain was removed so that participants could see the loudspeakers.

The first two series of experiments all applied HRTFs that were measured with a mannequin wearing the headphones used in the experiment. This ensured that the synthesis was technically as close as possible to its real-world counterpart. Comb filtering and direction-dependent coloration effects, which are introduced through the open headphones, are present in both synthesis and real-world presentation. It is debatable whether this exact modeling positively affects the plausibility ratings or whether it does not match the participants' inner reference and even negatively affects the results.

Moreover, in a consumer application, it could be interesting to apply simplified HRTF sets to save computation and memory. Ackermann et al. [17] showed that even a simplified spherical head model, the motion tracked binaural (MTB) synthesis approach, which only synthesizes interaural time difference and interaural level difference (ILD), can achieve plausible results. Hence, the next variation of experiments compares two sets of HRTFs, one set measured with the Neumann KU100, which did not wear the AKG K1000, and one set simulated with a spherical head model. Hence, the auralization systematically deviates from the ideal model of the real sound source by degrading the accuracy of the HRTF set. The hypothesis is that the perceived plausibility decreases notably for the spherical head model.

Classical research on binaural technology focused on dynamic binaural synthesis with 3 DoF (i.e., adapting the sound field only to the listeners' head orientation). However, actual research is increasingly shifting toward synthesis with 6 DoF, allowing listeners to navigate within the auralized space. The last iteration investigated the influence of 6 DoF tracking and determined whether, in a sitting scenario, minor head movements require consideration in binaural synthesis. Throughout all experiments, the incidence direction, loudspeaker transfer function, and distance attenuation were updated according to the listener's head orientation and position. Although the participants

remained seated during the experiments, they constantly made subtle translational head movements. To examine if this affects the plausibility of the experience, the same experiment was performed while updating only the binaural synthesis to the listener's head orientation.

Some of the results from this study have been previously presented in [31] and [32]. In summary, this work provides a comprehensive overview of the five experiments investigating various factors affecting the plausibility of binaural rendering in AAR systems. The work comprises the following experiments:

1. Head-related impulse response KU100 with K1000 baseline,
2. Influence of SH order,
3. Influence of visual cues,
4. Influence of different head-related impulse response sets, and
5. Influence of small translational head movements.

1 METHODS

1.1 Setup: Parametric Binaural Renderer

The (position) dynamic binaural rendering was performed with Max/MSP (Version 8.1.5). It is a slightly adapted version of the rendering presented in [32]. The listener's head orientation and position and the position of the loudspeakers is tracked with an Optitrack Flex 13 system with the Motive Software (Version 2.2.0). Via Open Sound Control, the tracking data are transferred to a Python middleware and forwarded to the Max software.

The direct sound is rendered by convolving the input signal with a pair of HRTFs corresponding to the relative direction of listener's head orientation and the position of the loudspeaker. Further, the direct sound is attenuated according to the $\frac{1}{r}$ distance law and filtered with the loudspeaker frequency response for the respective relative angle of the loudspeaker and listener position. The convolution, fading, and interpolation is performed with the spat.binaural~module from the Max spat5 library [33], which is publicly available.[1] It can read arbitrary HRTF sets in Spatially Oriented Format for Acoustics [34, 35] and provides several options for fading, interpolation, and ILD modifications. The default option of all-pass interpolation together with a linear fading over 5 ms was applied.

The frequency responses of the loudspeaker were measured in the anechoic chamber at TH Köln for 2,702 points on an $N = 44$ Lebedev grid. They were then resampled in the SH domain to obtain a horizontal set of 360 directions in 1° steps. Hence, during the reproduction, the loudspeaker directivity was updated only with respect to the relative horizontal angle, which is a fair assumption since the participants sat on a chair during the experiment with their head approximately on the same height as the acoustical center of the loudspeakers. The directivity was further octave-band smoothed and exported as 2,048 tabs minimum phase finite impulse response filters.

The renderer is generally capable of synthesizing a full 6 DoF VAE with reverberation and early reflections based on the method described by [6, 36]. However, in this work, only the direct sound part was used (i.e., convolution with HRTFs and loudspeaker transfer function, as well as distance variation).

The digital-to-analog converter was an RME UFX 2 interface with a sampling rate of 48 kHz and a buffer size of 256 samples. All experiments were performed with the extra-aural AKG K1000 headphones in the open configuration to reduce the shadowing effect of the headphones during the playback of the real sound sources. The real sound sources were presented with Genelec 1029a loudspeakers.

All loudspeakers were placed with a distance of 1.62 m to the listener at azimuth directions $\phi = 90°$,[2] $\phi = 0°$, $\phi = 270°$, and $\phi = 225°$ and arranged such that they all faced the listener. Although the participants remained seated during the experiment, they were allowed to make translational head movements. Accordingly, the loudspeaker directivity and direct sound level were updated. Due to the loudspeaker distance of 1.62 m, any near-field effects such as ILD modifications could be neglected [37].

1.2 Procedure

All tests were conducted in the anechoic chamber at TH Köln. The participants sat on a swivel chair and were provided with a keyboard. After participants put on the headphones, adjusted the wearing position, and calibrated the tracking system, the headphones were fixed with a strap to prevent any repositioning during the test. At the beginning of each experiment, a guitar sample was played back from the frontal speaker, to ensure that all participants had a similar expectation of the real acoustic event in an anechoic chamber. All participants were informed that the sample was coming from a real loudspeaker. This test sample was not played back during the actual experiment.

After the training stimulus, 100 trials were played back either via real loudspeaker or virtually through the headphones. The number of virtual and real sources was slightly unbalanced (i.e., 47 and 53, 48 and 52, or 49 and 51) to avoid participants making assumptions about the number of correct and false ratings. It was also ensured that a real source was never followed by a virtual one, or vice versa, in the same position. After each trial, the participants had to answer the question: "Did the sound come from one of the loudspeakers?" by pressing a key labeled "yes" or "no" on the keyboard. The subjects took between 10 and 15 min to complete all the trials.

1.3 Data Analysis

A standard method for analyzing the results of a plausibility test is Signal Detection Theory (SDT) [38]. SDT decomposes participants' response behavior into two mea-

[1] https://forum.ircam.fr/projects/detail/spat/.

[2] ϕ denotes the horizontal angle ranging from 0° to 359°, and θ denotes the colatitude angle ranging from 0° to 180°.

sures: sensitivity d and bias β. Sensitivity is a criterion-free measure of the perceived sensory difference between the two conditions and describes how well participants can discriminate between virtual and real reproduction. The bias indicates whether the participant tends to believe in the realness of the sound source or a tendency to vote for the virtual. To estimate sensitivity and bias, SDT considers four possible outcomes of the yes/no task: hit (virtual reproduction is correctly reported), false alarm (a real source has been reported as virtual), miss (a virtual source has been reported as real), or correct rejection (a real source has been reported as real). The rate of hits (p_{hit}) and false alarms (p_{fa}) fully describe the participant's performance. With the inverse cumulative normal distribution $Z(p)$ the sensitivity can be estimated with

$$d' = Z(p_{\text{hit}}) - Z(p_{\text{fa}}). \tag{1}$$

An estimated sensitivity of $d' = 0$ indicates no difference between the virtual and real presentation and therefore indicates perfect plausibility. $d' \neq 0$ indicates that participants were able to determine whether the presentation was real or virtual. It is often easier to interpret the difference in terms of forced-choice values. With the underlying assumption for SDT of an equal-variance Gaussian distributed observer model d can be converted to two alternative forced choice probablities (P2AFC) with

$$p_{\text{2AFC}} = \Psi(d/p), \tag{2}$$

with $\Psi()$ being the cumulative standard normal distribution. The bias can be calculated in different ways; in this paper, it is calculated it with

$$\beta' = e^{-\frac{Z(p_{\text{hit}})^2}{2} + \frac{Z(p_{\text{fa}})^2}{2}}, \tag{3}$$

according to [38], which is robust to rounding errors. A $\beta' > 1$ can be interpreted as indicating that participants tend to believe in the realness of the source, and with a $\beta' < 1$, they tend to believe in a simulated source.

The statistical analysis in this work is based on the individual d' for each participant. Alternatively, Neidhardt and Zerlik [18], for example, discussed the use of a pooled d', which is calculated from the pooled p_{hit} and p_{fa} of all participants. To make a statement about whether a system can be considered plausible, Lindau and Weinzierl proposed to test whether the individual d' is significantly different from a critical d_{min} using null hypothesis testing. Perfect plausibility would mean no significant difference from d_{min}. However, this threshold is arbitrarily chosen and is a point of debate in the literature. Lindau and Weinzierl suggested a $d_{\text{min}} = 0.1777$, which corresponds to $p_{\text{2AFC}} = 55\%$ and thus allows 5% difference from a pure guessing rate of 50% in an equivalent forced-choice task. Other studies use a less strict criterion and refer to a $p_{\text{2AFC}} = 75\%$, which corresponds to the point of subjective equality on the psychometric function [16]. This study refers to the limit of $d_{\text{min}} = 0.1777$ to maintain comparability across the majority of existing studies.

Another limitation of analyzing the SDT outcome with hypothesis testing is that a nonsignificant difference from d_{min} not necessarily indicates the absence of a difference between the two auralizations. For this reason, this work also reports the Bayes factor (BF; Jeffreys-Zellner-Siow scaling factor 0.707), which can be interpreted as the probability of the null hypothesis of the t tests [39].

A related issue is the estimation of an optimal sample size. Generally, a direct proof of the null hypothesis (perfect plausibility, $d' = 0$) is impossible. As derived in Lindau and Weinzierl [14], an alternative approach is to reject a minimal detectable effect (minimum effect hypothesis). A formulation of a meaningful minimum effect can be done in terms of 2AFC values, which, as a reminder, can be converted to d' using Eq. (2). The calculation of the required sample size depends on the variance of d' and is influenced by hit and false alarm rates as well as assumptions about noise and signal conditions. These calculations determine the required number of trials to achieve acceptable error levels in hypothesis testing. This work defines a $P_{\text{2AFC, min}} = 55\%$ and assumes an equal number of real and virtual stimuli. Allowing type I and type II error rates of 0.25 and 0.005, respectively, rejecting a $P_{\text{2AFC, min}} = 0.55\%$ requires 1,071 decision ratings. With 100 trials per experiment (50 noise and 50 signal), this equates to approximately 11 participants per experiment. It should be noted that this is the minimum number of participants for $d' < 1$.

Other ways of analyzing the results of the yes/no paradigm are summarized, for example, in Neidhardt and Zerlik [18]. The SDT analysis was performed using the dprime method from the R psycho package (version 0.6.1). Bayesian t tests were calculated using the BayesFactor package, according to Morey and Rouder [39].

1.4 Experiment 1: Baseline

1.4.1 Stimuli

The initial baseline experiment applies a binaural synthesis based on HRTFs of a Neumann KU100 mannequin head wearing the AKG K1000 headphone during the measurement [40] since this represents a synthesis technically closest to the playback of the real loudspeakers. In this case, comb filtering or coloration effects due to the open AKG K1000 headphones are present in both the real and virtual reproductions. Furthermore, in this experiment, all loudspeakers were covered with an acoustically transparent curtain. This ensured that participants only focused on the auditory cue. The entire binaural chain was equalized with headphone compensation filters calculated according to the procedure described by Erbes et al. [41]. As anechoic dry audio signals, 25 different samples were used (three drum, three piano, five guitar, one saxophone, one cello, one harmonica, one clarinet, one oboe, one violin, one owl sound, two female and two male speech, and one female and two male singing samples). Since the anechoic chamber at TH Köln cannot be considered anechoic below 200 Hz, all test signals were high-pass filtered with a cut-off at 200 Hz, to ensure equal conditions for virtual and real presentation. These audio signals were used for all experiments in this work.

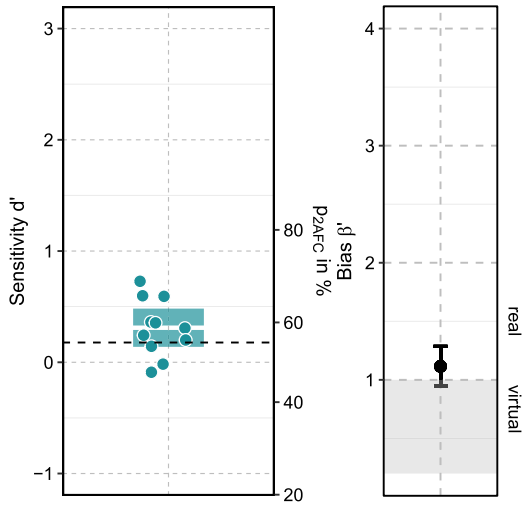


Fig. 1. Experiment 1: Individual d', average $\bar{d}'$, and corresponding 95% CIs, as well as average bias $\bar{\beta}'$ and error bars.

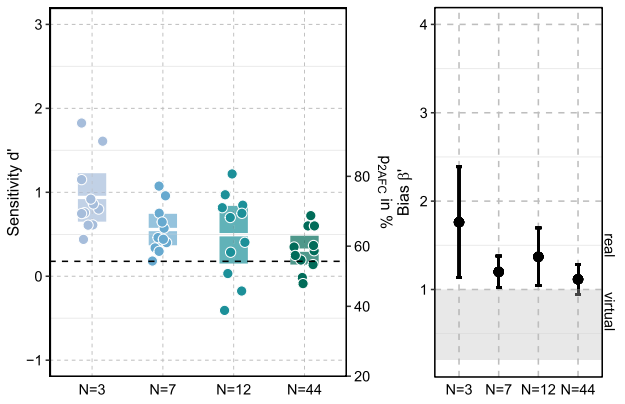


Fig. 2. Experiment 2: Individual d', average $\bar{d}'$, and corresponding 95% CIs, as well as average bias $\bar{\beta}'$ and error bars for rendering based $N = 3$, 7, 12, and 44.

1.4.2 Participants

According to the sample size estimation of Lindau and Weinzierl [14], which was also adopted by Pike et al. [15], between 11 and 13 participants were recruited for each listening test. In Experiment 1, 11 participants took part. Although half of them already listened to dynamic binaural synthesis, none of them had participated in a plausibility test before, which is why they will be considered naive listeners in the following.

1.4.3 Results

A first overview of the results is presented in Fig. 1. The figure on the left shows the individual d' for each participant as dots, the average d' ($\bar{d}'$) as the white bar, and the 95% confidence intervals (CIs) around the mean as boxes. The dashed line indicates d_{min}. The plot on the right depicts the average $\bar{\beta}'$ as a black dot, and the 95% CIs as error bars. The gray area below 1 indicates the area for the decision tendency toward virtual. It can be observed that $\bar{d}'$ is slightly above the d_{min} line ($\bar{d}' = 0.31$).

A Lilliefors test for normality of the individual d' revealed no violations of the assumption of normality distribution ($p = 0.761$), which is why parametric tests were applied. A (one-sided) one-sample t test indicated no significant difference to the $d_{\text{min}} = 0.1777$ [$t(10) = 1.73$, $p = 0.058$, $d = 0.52$]. The Bayes factor $\text{BF}_{01} = 1.084$ indicates a weak evidence for the null hypothesis [42].

1.5 Experiment 2: Varying Spatial Resolution

1.5.1 Stimuli

The second experiment evaluates the influence of spatial resolution on the plausibilty of binaural synthesis. The spatial resolution is scaled by varying the SH order. The employed HRTF set, which was also used in Experiment 1, was measured on a 2,702 sampling point Lebedev grid, which can be considered the full spatial resolution. For generating HRTFs sets of reduced spatial order, the reference HRTFs set was transformed at the maximum order of $N = 44$ using the spatial Fourier transform. The SH coefficients were then truncated at the orders $N = 3$, $N = 7$, and $N = 12$ and inverse transformed to the same 2,702 sampling point Lebedev grid. The spectral roll-off caused by SH truncation was diffuse-field equalized according to Ben-Hur et al. [43].

1.5.2 Participants

A total of 44 participants (11 for each SH order) took part in the experiments, whereby the results from Experiment 1 for the $N = 44$ condition was the same as for Experiment 2. Hence, the participants for the experiments for the $N = 3$, $N = 7$, and $N = 12$ conditions were not the same as for the $N = 44$ case. Again, half of the participants had no experience with binaural technology; none of them had participated in a plausibility test before.

1.5.3 Results

An overview of the individual d' is depicted in Fig. 2. It can be observed that the mean $\bar{d}'$ increases with decreasing order, suggesting that lower SH orders lead to less plausible auralization. All means of the bias lay within the real tendency area.

A Lilliefors test showed no violations of the assumption of normality distribution for all orders. Experiment 1 showed that the t tests against d_{min} yield no significant difference for $N = 44$. However, for $N = 12$ [$t(10) = 2.06$, $p = 0.033$, $d = 0.62$, $\text{BF}_{01} = 0.729$, $\bar{d}' = 0.49$], $N = 7$ [$t(10) = 4.51$, $p = 0.001$, $d = 1.35$, $\text{BF}_{01} = 0.027$, $\bar{d}' = 0.56$], and $N = 3$ [$t(10) = 5.88$, $p < 0.001$, $d = 0.91$, $\text{BF}_{01} = 0.005$, $\bar{d}' = 0.938$], the t tests indicated a significant difference from d_{min}. Paired t tests between different orders yield significant differences between $N = 44$ and $N = 3$ [$t(10) = 4.17$, $p < 0.001$, $d = 1.77$], $N = 44$ and $N = 12$ [$t(10) = 2.48$, $p = 0.024$, $d = 1.06$], and $N = 44$ and $N = 7$ [$t(10) = 2.21$, $p = 0.04$, $d = 1.77$].

Some participants reported they might have changed their rating strategy throughout the experiment. To analyze this, the 100 trials were divided into four epochs, with 25 trials

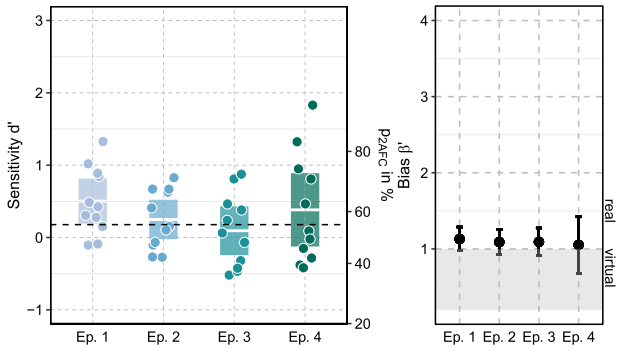


Fig. 3. Experiment 2: Individual d', average $\bar{d}'$, and corresponding 95% CIs, as well as average bias $\bar{\beta}'$ and error bars with respect for epochs, for the $N = 44$ condition.

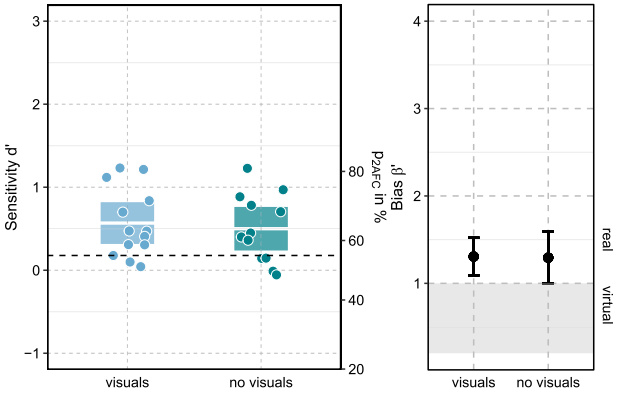


Fig. 4. Experiment 3: Individual d', average $\bar{d}'$, and corresponding 95% CIs, as well as error bars of the bias $\bar{\beta}'$ for condition with visual loudspeakers (visuals) and without (no visuals).

each. Fig. 3 depicts the mean $\bar{d}'$ and $\bar{\beta}'$ with respect to the four epochs for $N = 44$. Interestingly, the $\bar{d}'$ decreases over time; however, in the last epoch, it notably increases again. For sake of completeness, plots separate for epochs for all other experiments are provided in the supplementary material. This article will only show and discuss those with interesting observations.

1.6 Experiment 3: Influence of Visual Cues

1.6.1 Stimuli

To examine if the visual cue affects the perceived plausibility of binaural rendering, the baseline experiment was repeated but without the curtain. The condition with curtain is denoted as "no-visual," and the condition with visible loudspeakers is the "visual" condition in the remainder.

1.6.2 Participants

From Experiment 3 onward, not only naive listeners took part in the experiments. Each participant who had already taken part in one or more plausibility tests (including pilot tests) is considered an expert listener from now on. To ensure equal distribution of naive and expert listeners, the no-visual experiment was repeated; hence, the results are not the same as in Experiments 1 and 2. Twelve participants took part in the no-visual experiment (six naive and six expert listeners), and 13 participants took part in the visual experiment (seven naive and six expert listeners).

1.6.3 Results

Fig. 4 shows the individual d' for all participants separately for the visual and no-visual conditions. It can be observed that $\bar{d}'$ of the visual case is slightly above that of the no-visual. Both are notably higher than the d_{min} line. Fig. 5 shows the results separately for naive and expert listeners. For both conditions, the means for the naive listeners are slightly lower than for the experts.

The Lilliefors test indicates that all results are normal distributed. The t test against d_{min} for all participants yields a significant difference for both visual condition [$t(12) = 3.38, p = 0.003, d = 0.94, \text{BF}_{01} = 0.105, \bar{d}' = 0.57$] and no-visual condition [$t(11) = 2.705, p = 0.01, d = 0.78,$

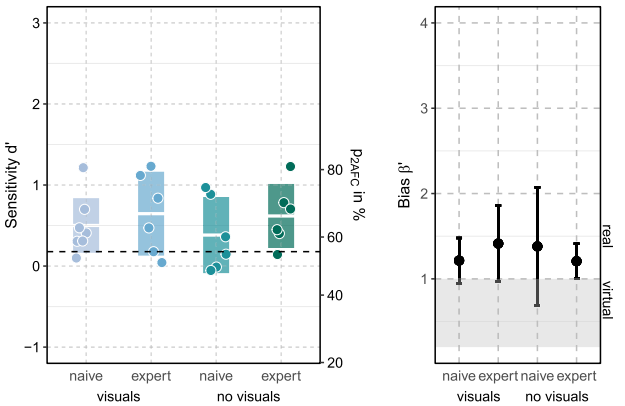


Fig. 5. Experiment 3: Individual d', average $\bar{d}'$, and corresponding 95% CIs, as well as error bars of the bias $\bar{\beta}'$ for condition with visual loudspeakers (visuals) and without (no visuals) with respect to naive and expert listeners.

$\text{BF}_{01} = 0.3, \bar{d}' = 0.5$]. Both BFs indicate a moderate probability for the alternative hypothesis (i.e., there is a slight difference between real and virtual reproduction). Moreover, in the visual test, there was a significant difference for the naive listeners [$t(6) = 2.35, p = 0.029, d = 0.89, \text{BF}_{01} = 0.555$] and expert listeners [$t(5) = 2.33, p = 0.034, d = 0.95, \text{BF}_{01} = 0.594$]. For the no-visual condition, only the data of expert listeners [$t(5) = 2.86, p = 0.018, d = 1.17, \text{BF}_{01} = 0.371$] showed a significant difference. Pairwise t tests between visual and no-visual conditions did not reveal any significant difference. Figs. 6 and 7 show the $\bar{d}'$ and $\bar{\beta}'$ with respect to epochs. For the no-visual condition, it can be observed that whereas the $\bar{d}'$ for the first epoch is relatively small, it abruptly increases after the first epoch. For the visual condition, it first decreases after the first epoch but then successively increases over the last two epochs.

1.7 Experiment 4: Influence of Head-Related Transfer Functions

1.7.1 Stimuli

From Experiment 4 onward, the loudspeakers were not hidden behind a curtain anymore, such that participants

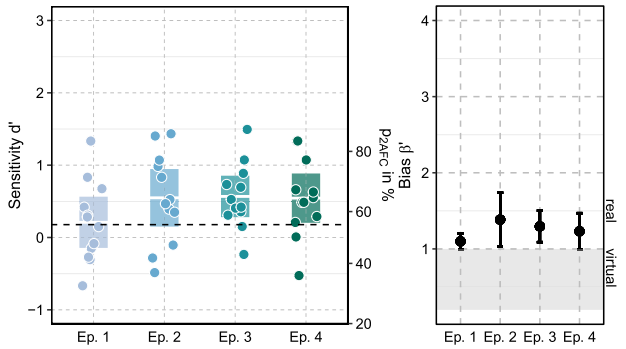


Fig. 6. Experiment 3: Individual d', average $\bar{d}'$, and corresponding 95% CIs, as well as error bars of the bias $\bar{\beta}'$ for the condition without visual loudspeakers (no visuals) with respect to time epoch.

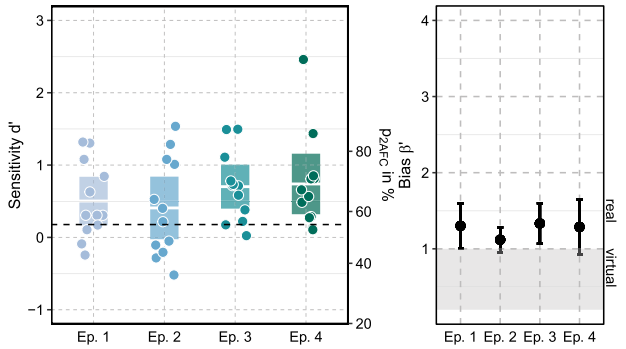


Fig. 7. Experiment 3: Individual d', average $\bar{d}'$, and corresponding 95% CIs, as well as error bars of the bias $\bar{\beta}'$ for the condition with visual loudspeakers (visuals) with respect to time epoch.

could always see the loudspeakers. Hence, the results of the following experiments will always be compared to the visual results from Experiment 3. Experiment 4 examined two more HRTF sets: a Neumann KU100 HRTF set [44] without AKG K1000, denoted as clean KU100, and an HRTF set simulated according to a spherical head model (SHM) with a radius of 0.0875 m.

1.7.2 Participants

Thirteen participants took part in the clean KU100 experiment (seven naive and six expert listeners). 13 participated in the SHM experiment (seven naive and six expert listeners).

1.7.3 Results

Two observations can be made from Fig. 8: The $\bar{d}'$ values of the clean KU100 are notably smaller than the $\bar{d}'$ from the KU100 with AKG K1000. The $\bar{d}'$ values of the SHM are in the same range as that of the KU100 with AKG K1000, however, with higher variances.

Fig. 9 depicts the results separately for naive and expert listeners. Interestingly, only for the AKG K1000, the $\bar{d}'$ of the naive listeners is below that of the experts. The Liliefors test showed that only the results of the SHM are not normally distributed. However, the same parametric tests as in all other experiments were performed. The results

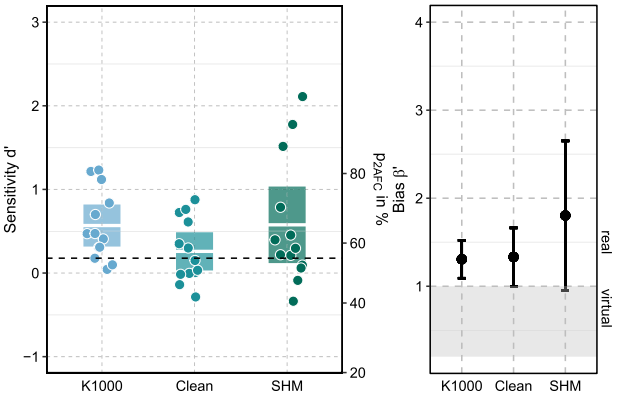


Fig. 8. Experiment 4: Individual d', average $\bar{d}'$, and corresponding 95% CIs, as well as error bars of the bias $\bar{\beta}'$ for the three different HRTF sets KU100 with AKG K1000, clean KU100, and HRTFs based on the spherical head model.

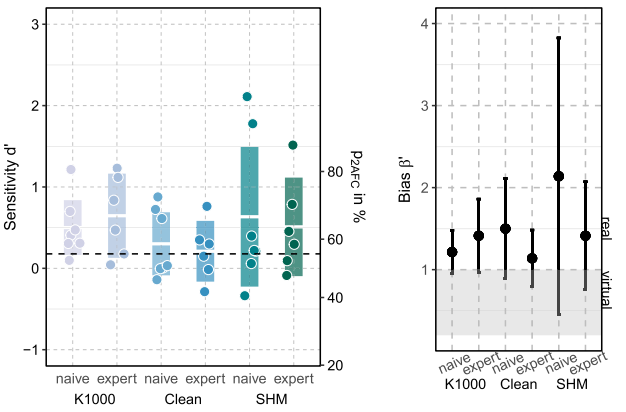


Fig. 9. Experiment 4: Individual d', average d', and corresponding 95% CIs, as well as error bars of the bias $\bar{\beta}'$ for the three different HRTF sets separately for naive and expert listeners.

for the clean KU100 HRTFs are not significantly different from d_{min} for all participants [$t(12) = 0.78$, $p = 0.226$, $d = 0.22$, $\text{BF}_{01} = 2.774$; $\bar{d}' = 0.26$], naive participants [$t(6) = 0.78$, $p = 0.232$, $d = 0.3$, $\text{BF}_{01} = 2.216$], or expert listeners [$t(5) = 0.22$, $p = 0.417$, $d = 0.09$, $\text{BF}_{01} = 2.625$], where all BF_{01} values can be interpreted as a weak tendency for the null hypothesis. Also, for the SHM, no significant differences from d_{min} were found for all listeners [$t(12) = 1.9$, $p = 0.041$, $d = 0.53$, $\text{BF}_{01} = 0.9$, $\bar{d}' = 0.26$], naive [$t(6) = 1.3$, $p = 0.121$, $d = 0.49$, $\text{BF}_{01} = 1.513$], and expert [$t(5) = 1.41$, $p = 0.109$, $d = 0.58$, $\text{BF}_{01} = 1.336$]. Paired t tests between all conditions revealed a significant difference between clean KU100 and KU100 with K1000 HRTFs [$t(23) = 1.98$, $p = 0.03$, $d = 0.78$], but not for the SHM case.

Fig. 7 depicts the mean $\bar{d}'$ and $\bar{\beta}'$ with respect to the four epochs for the KU100 with K1000. It can be observed that the $\bar{d}'$ increases over time epoch, showing that the hit rate improves slightly. The values increase to such an extent that the CI no longer overlaps d_{min} from epoch 3 onward. In all epochs, there is still a tendency toward the real. An opposite trend can be observed for the clean KU100

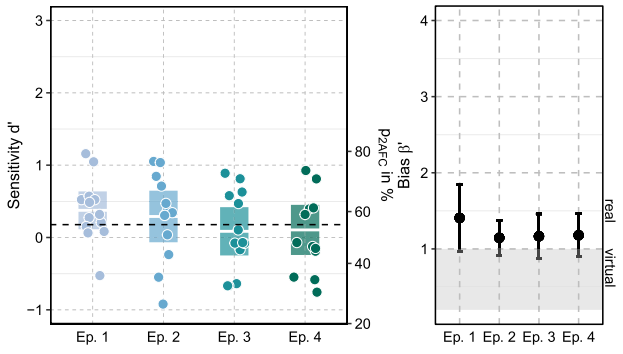


Fig. 10. Experiment 4: Individual d', average $\bar{d}'$, and corresponding 95% CIs, as well as error bars of the bias $\bar{\beta}'$ for the clean KU100 HRTFs with respect to time epoch.

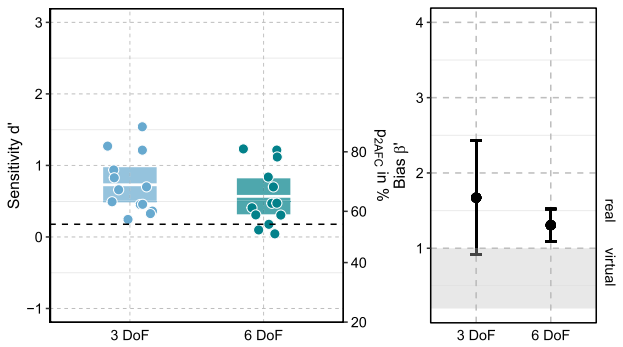


Fig. 11. Experiment 5: Individual d', average $\bar{d}'$, and corresponding 95% CIs, as well as error bars of the bias $\bar{\beta}'$ for the 3 DoF and 6 DOF conditions.

(Fig. 10), showing a decreasing trend for $\bar{d}'$ over the time. After epoch 2, $\bar{d}'$ is even below d_{min}. The bias $\bar{\beta}'$ shows that the decision tendency drifts slightly toward the virtual from epoch 2 onward.

1.8 Experiment 5: Translational Head Movements

1.8.1 Stimuli

The last experiment evaluates how small translational head movements affect plausibility. So far, all binaural renderings account for the listener's head movement with 6 DoF. To test whether 6 DoF improves plausibility in a seated scenario, the same experiment was conducted with the same instructions to the participant, but the rendering was updated with only 3 DoF.

1.8.2 Participants

Thirteen participants took part in the experiment; seven of them were naive listeners, and six of them were experienced listeners.

1.8.3 Results

From Fig. 11, it can be observed that $\bar{d}'$ of the 3 DoF condition is slightly higher than that of the 6 DoF condition. For the 3 DoF case, notable higher variances can be seen. It can thus be concluded that 3 DoF rendering breaks

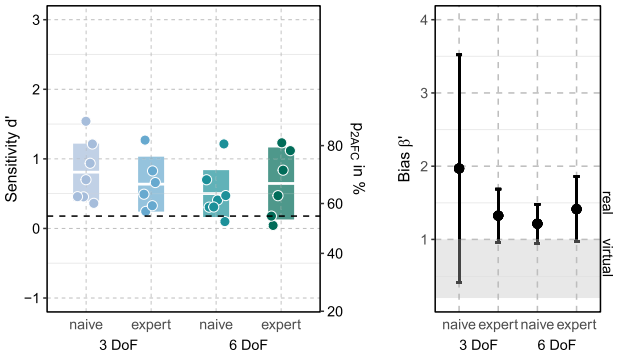


Fig. 12. Experiment 5: Individual d', average $\bar{d}'$, and corresponding 95% CIs, as well as error bars of the bias $\bar{\beta}'$ for the 3 DoF and 6 DoF condition with respect to naive and expert listeners.

plausibility for some participants and not others. Fig. 12 shows that for the 3 DoF case, the $\bar{d}'$ of the experts is lower, and for the 6 DoF case, experts achieve slightly higher $\bar{d}'$.

Both conditions are normally distributed according to the Lilliefors test. For both 3 DoF [$t(12) = 4.9$, $p < 0.001$, $d = 1.36$, $\text{BF}_{01} = 0.011$, $\bar{d}' = 0.73$] and 6 DoF [$t(12) = 3.38$, $p = 0.005$, $d = 0.94$, $\text{BF}_{01} = 0.105$, $\bar{d}' = 0.57$] for all participants, there was a significant difference to d_{min}. Both values for BF_{01} can be interpreted as a moderate probability for the alternative hypothesis. Also, for the 3 DoF case for the naive listeners [$t(6) = 3.77$, $p = 0.005$, $d = 1.42$] and expert listeners [$t(5) = 3$, $p = 0.03$, $d = 1.22$], as well as for the naive [$t(6) = 2.35$, $p = 0.029$, $d = 0.89$] and expert [$t(5) = 2.33$, $p = 0.034$, $d = 0.95$] listeners of the 6 DoF case, there were significant differences. The t tests did not reveal a significant difference between 3 DoF and 6 DoF for all participants [$t(12) = 1.0$, $p = 0.164$, $d = 0.39$].

Analysis of the tracking data only in the horizontal plane did not yield any systematic difference in the ratings for the 3 DoF or 6 DoF case. For this, the average of the absolute values in x and y direction across all participants and all trials were calculated, and maximum movement across all participants and trials was calculated. For the 6 DoF condition, participants moved on average with 8.95 cm (max: 33.71 cm) in x and 4.32 cm (max: 23.09 cm) in y direction, which corresponds to an angular change of 2.12° (max: 9.56) and distance variation of 7.51 cm (max: 33.83 cm) for the respective sound sources. For the 3 DoF condition, participants moved on average with 9.89 cm (max: 22.11 cm) in x and 6.23 cm (max: 21.58 cm) in y direction, which corresponds to an angular change of 2.52° (max: 7.75°) and distance variation of 9.04 cm (max: 22.11 cm) for the respective sound sources. Interestingly, although an angular change of more then 2° is in the range of the minimum audible angle [45], it does not significantly affect the plausibility. The level change corresponding to the distance variation of about 20 cm is below the JND. Corresponding plots, as well as the tracking data for all other experiments, are provided along with supplementary materials.[3]

[3]Supplementary materials for the review process are available at https://doi.org/10.5281/zenodo.13319867.

2 DISCUSSION

The results indicate that for the majority of the cases, the estimated average $\bar{d}'$ falls within a range where participants could not reliably distinguish between real and virtual reproduction. This aligns with verbal reports following each experiment, which suggest that in the majority of cases, participants were just guessing.

2.1 Experiment 1

Experiment 1, which serves as a baseline experiment shows that the system yields plausible results, even in the strict sense of hypothesis testing against d_{min}.

2.2 Experiment 2

Experiment 2 shows that the SH order clearly affects plausibility. This is generally not surprising, since lower SH orders lead to less localization and a blurry image of the sound source. Moreover, the SH orders, which result in d' values in the range at which the system can be considered plausible, are way below the thresholds of indistinguishability reported for the direct sound component (e.g., in Lübeck et al. [22]). Assuming a similar trend as for the indistinguishabilty thresholds that the addition of room acoustics reduces the required SH order for the direct sound, a few SH orders below $N = 12$ are certainly enough for plausible rendering in consumer applications. This is a promising piece of information since it is in the order of magnitude of which can be implemented with current microphone arrays and calculated in current real-time applications. Moreover, the SH transform was performed without any HRTF preprocessing, such as time alignment [46] or the magnitude-least squares optimization [47, 19], which lead to significant perceptual improvements. It can be assumed that such algorithms also have an effect on plausibility, and even lower SH orders can result in plausible auralization.

2.3 Experiment 3

Experiment 3 shows that the visual cue only slightly affects plausibility. Visual inspection showed a weak tendency that the absence of visual cues improve the plausibility. However, only for naive listeners it can be observed that the d' for the no-visual condition are not significantly different from d_{min}. This is surprising since it could be argued that nonindividual HRTFs increase the localization error, which in the case of visible loudspeakers might reduce the perceived plausibility. The findings drawn from the present results consistent with the findings reported by Riedel et al. [30] that for nonindividual HRTFs, no significant differences in the horizontal localization was observed for visual and no-visual conditions. In the present study, all loudspeakers were in the horizontal plane as well. Another explanation for the weak influence of the visibility of loudspeakers could be that the ventriloquist effect causes participants to believe that the sound originated from the loudspeakers, although the auditory system localizes the sound from a slightly different position. It can thus be concluded that the plausibility results are influenced by different and potentially even contrasting perceptual effects. The plots of the mean $\bar{d}'$ with respect to epochs (Figs. 6 and 7) indicate the interesting trend that for the no-visual condition, participants were unsure at the beginning of the test but improved their rate of correct ratings successively over the course of the experiments. For the visual condition, the $\bar{d}'$ increased abruptly after the first epoch. Hence, a slight learning effect can be observed for both conditions, which, however, started notably faster when participants could see the loudspeakers. Bailey and Fazenda [27] reported that in a fully virtual visual environment, adding visual sound sources increases plausibility.

2.4 Experiment 4

Experiment 4 shows that the HRTFs of the KU100 without the AKG K1000 lead to smaller $\bar{d}'$ compared to the K1000 condition. In fact, they were the smallest $\bar{d}'$ across all experiments. Independent of the condition, some of the participants reported perceiving a difference from the inner reference when the sound source came from behind, a finding that has been similarly observed by Neidhardt and Zerlik [18]. This suggests that the synthesis that is physically closer to the real sound source does not necessarily result in a smaller d'. Some participants indicated that they became more uncertain during the test or assumed that the majority could not be real and changed their response behavior. This indicates that the sensitivity changes throughout the trials and is therefore reflected in the epoch results. In the case of the HRTF sets with the AKG K1000 headphones, it can be seen that the sensitivity increases slightly over the epochs. This suggests that the participants were able to detect cues in the HRTFs over the epochs and use them as a guideline, resulting in a slight learning effect. Interestingly, this is not the case with the clean KU100 HRTFs, where the mean $\bar{d}'$ decreases over the epochs. This is in line with the participants' statements that they became more uncertain throughout the trials and could no longer reasonably discriminate between real and virtual sources. The mean $\bar{d}'$ of the SHM HRTFs is comparable to that of the K1000. However, because of the high variance, the t tests did not reveal a significant difference between the two. For some participants, there were d' values that were considerably smaller than d_{min}, while for others, the values were above 1.5. This suggests that for some participants, the SHM model performed well, while for others, it was relatively straightforward to distinguish between simulation and real sound sources. Ackermann et al. [17] evaluated the plausibility of binaural rendering with the MTB approach using an eight-channel and a 16-channel spherical microphone array. In contrast to the present study, Ackermann et al.'s study was conducted in a reverberant environment (reverberation time$_{60} = 0.28$ s). Since no HRTF processing is involved and only interaural time differences and ILDs based on a spherical head model can be reproduced, this model is comparable to the present spherical head model that uses the same radius as the MTB. Consequently, the reported d' (eight-channel $d' = 0.62$, 16-channel $d' = 0.53$)

are in a similar range to that of these results. Also in Ackermann et al. [17], notable high variances have been observed.

2.5 Experiment 5

Experiment 5 indicates that in a seating scenario, adapting the sound field only with 3 DoF only slightly affects the plausibility of binaural rendering. Even though t tests did not reveal a significant difference, the distribution of the individual d' suggests that the d' values of the 6 DoF condition are slightly lower than that of the 3 DoF case.

2.6 General Discussion

Conducting the plausibility experiments in an anechoic environment comes along with one limitation: It is unclear what listeners expect from a real sound source in the anechoic chamber. The concept of plausibility is based on this expectation of the real sound source, which is referred to as the inner reference. The majority of people are not used to anechoic conditions, which results in a lack of a strong inner reference. It can be argued that both virtual and real reproduction are, to some extent, perceived as artificial. To address this, at the beginning of each experiment, a single training sample was presented by a real loudspeaker to establish an inner reference. It is unclear if a single sample can fully establish the inner reference; however, a longer training period may have resulted in a learning effect, potentially influencing the plausibility tests. On the other hand, starting with Experiment 3, participants were recruited who had already completed a plausibility test in anechoic conditions and thus have established an inner reference, which, however, turned out to have no significant influence.

As discussed in SEC. 1.3, it is questionable if hypothesis testing against a strict and generally arbitrarily chosen limit of d_{min} is a reasonable approach for determining the plausibility of binaural rendering and especially investigating small changes in the rendering system. Generally, it can be observed that rigorous application of hypothesis testing against $d_{\text{min}} = 0.1777$ does not align with the informal feedback of the majority of participants. Most of them indicated that they were just guessing, although this is not reflected in the results of the hypothesis tests. Moreover, it can be observed that one system is perceived as being plausible by some participants, while it was not plausible for others, as for the SHM model, for example. This finding is also not represented by reporting the results of a hypothesis test. d_{min}

On the other hand, establishing a unified definition of a plausibility criterion would help to compare different rendering systems. A last problem is that a nonsignificant difference from is not clear evidence for the absence of an effect. Consequently, in the sense of hypothesis testing, one can only conclude that a system is not plausible. Furthermore, hypothesis testing reduces plausibility to a binary concept, whereby a system can only be classified as either plausible or not plausible. However, it may be useful to compare systems based on a continuous scale, which allows saying one system is more plausible than another.

Here is a main contribution of the present work, which not only reports absolute estimated d' but rather shows how the change of critical factors of binaural rendering affects d'. Overall, the presented results indicate that even large manipulations of key aspects of binaural rendering have remarkably little impact on the sensitivity d. Table 2 provides an overview of d' values reported in existing literature. It covers exclusively the results of studies that strictly applied the paradigm introduced by Lindau and Weinzierl [14]. However, it is worth noting that despite the uniform adherence of the paradigm, there persist strong differences in the rendering across studies. Lindau and Weinzierl [14] employed HRTFs of an head-and torso simulator (HATS), which includes a model of the torso. This is also the case for the Kemar mannequin head used by Neidhardt and Zerlik [18]. Shoulder and torso reflections can affect sound source localization as shown, for example, by Gardner [48].

However, the extent to which this affects plausibility is questionable. Lindau and Weinzierl [14] and Pike et al. [15] used STAX SR-202 headphones for reproduction and also equalized the playback of the real loudspeakers accordingly. Since the STAX headphones are not fully open, it can be assumed that both real and virtual reproduction are significantly impaired. In particular, pinna cues important for spatial hearing that might affect the internal reference. Contrary, Neidhardt and Zerlik [18] and the present study used AKG K1000 headphones, which have much smaller shadowing effects. The results of the present study all have similar technical conditions, which helps to understand to what extent which parameters affect plausibility.

Besides the strict application of the procedure by Lindau and Weinzierl [14], other studies have investigated the plausibility with different methodologies. Neidhardt and Knoop [49], for example, assessed the plausibility of a 6 DoF dynamic binaural synthesis based on mannequin head measurements with a yes/no paradigm. After each trial, participants were asked to answer two questions: "Did you get the impression of walking toward/past a sound source?" and "Would you call this experience a plausible illusion of a sound source?" Brinkmann et al. [16] evaluated different binaural auralizations based on room simulation, and compared them with a mannequin head measurement in the real room. While showing a picture of the real room, they asked "Was this an audio example from a real room?" and analyzed the results with the SDT. They reported $d' = 0.07$, 0.3, and 0.75.

Arend et al. [9] investigated the influence of near-field cues on the perceived plausibility. The reference was a far-field HRTF set that was processed with distance variation functions. They compared it with HRTFs that where only intensity scaled. After each trial, participants had to choose the more plausible sound field representation based on their inner reference. Amengual Garí et al. [50] evaluated renderings based on the spatial decomposition method. They applied a test design in which two stimuli were presented from the same direction: either both stimuli were virtual or only one of them was virtual. This paradigm allows com-

Table 1. Estimated d' reported in literature. All studies strictly applied the plausibility test as proposed by Lindau and Weinzierl [14] and evaluated the results with SDT.

Study	Setting	Headphones	$\bar{d}'$
3 DoF HATS BRIRs [14]	Real/virtual	STAX SR-202	0.2956
3 DoF HATS BRIRs improved rendering [14]	Real/virtual	STAX SR-202	0.00512
3 DoF KU100 BRIRs [15]	Real/virtual	STAX SR-202	0.1954
6 DoF Kemar BRIRs [18]	Real/virtual	AKG K1000	1.05
3 DoF 8-channel MTB rendering [17]	Real/virtual	Sennheiser HD800S	0.62
3 DoF 16-channel MTB rendering [17]	Real/virtual	Sennheiser HD800S	0.53
6 DoF Ambisonics-based [27]	Real/virtual	Sennheiser HD800	3–10
3 DoF 3 different room simulations [16]	Measured/virtual	Sennheiser HD800	0.07, 0.3, 0.75

paring plausibility based on the participants inner reference or explicit reference. Most recently, McKenzie et al. [51] and Wirler et al. [52] introduced the concept of transfer-plausibility. Here, always two sound sources are presented, one of which is real and the other virtual. Participants need to indicate the virtual sound source.

3 CONCLUSION

This work presents results of five plausibility experiments in anechoic conditions, which systematically varied different aspects of binaural rendering. It is generally surprising to see that even large manipulations of important parameters of binaural rendering only weakly affect the plausibility. From the results, the following can be concluded:

- The spatial order affects plausibility; however, with orders above $N = 12$, plausible rendering can be achieved.
- Even though HRTFs measured with mannequin heads wearing the same headphones that are used for the playback are physically more correct, they do not improve plausibility.
- Making the sound sources visible only weakly affects plausibility.
- In a seated scenario, updating the sound field with 6 DoF only weakly affects plausibility.

Further research is recommended to extend the findings of the study and investigate the impact of room acoustics rendering on perceived plausibility. In a similar manner to the present study, parameters for room acoustics rendering could be systematically added and varied. This could include different numbers of early reflections or different accuracy of reverberation. Additionally, it would be valuable to examine how plausibility changes when presenting binaural synthesis in a scenario where listeners can walk.

4 ACKNOWLEDGMENT

Great thanks to Reality Labs Research at Meta for funding the project. Great thanks to all participants who took part in the listening tests.

5 REFERENCES

[1] A. Lindau, *Binaural Resynthesis of Acoustical Environments: Technology and Perceptual Evaluation*, Ph.D. thesis, Technische Universität Berlin, Berlin, Germany (2014 Jun.).

[2] A. Neidhardt and B. Reif, "Minimum BRIR Grid Resolution for Interactive Position Changes in Dynamic Binaural Synthesis," presented at the *148th Convention of the Audio Engineering Society* (2020 May), paper 10371.

[3] D. R. Begault, "Challenges to the Successful Implementation of 3-D Sound," *J. Audio Eng. Soc.*, vol. 39, no. 11, pp. 864–870 (1991 Nov.).

[4] V. Pulkki, "Spatial Sound Reproduction With Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516 (2007 Jun.).

[5] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial Decomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28 (2013 Jan.).

[6] C. Pörschmann, P. Stade, and J. M. Arend, "Binauralization of Omnidirectional Room Impulse Responses-Algorithm and Technical Evaluation," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx)*, pp. 345–352 (Edinburgh, UK) (2017 Sep.).

[7] A. Politis, S. Tervo, and V. Pulkki, "COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6802–6806 (Calgary, Canada) (2018 Apr.). https://doi.org/10.1109/ICASSP.2018.8462608.

[8] L. McCormack and S. Delikaris-Manias, "Parametric First-Order Ambisonic Decoding for Headphones Utilising the Cross-Pattern Coherence Algorithm," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 173–182 (Paris, France) (2019 Sep.). https://doi.org/10.25836/sasp.2019.26.

[9] J. M. Arend, S. V. Amengual Garí, C. Schissler, F. Klein, and P. W. Robinson, "Six-Degrees-of-Freedom Parametric Spatial Audio Based on One Monaural Room Impulse Response," *J. Audio Eng. Soc.*, vol. 69, no. 7/8, pp. 557–575 (2021 Jul.). https://doi.org/10.17743/jaes.2021.0009.

[10] K. Müller and F. Zotter, "Auralization Based on Multi-Perspective Ambisonic Room Impulse Responses,"

Acta Acust., vol. 4, no. 6, paper 25 (2020 Nov.). https://doi.org/10.1051/aacus/2020024.

[11] M. Slater, "Place Illusion and Plausibility Can Lead to Realistic Behaviour in Immersive Virtual Environments," *Philos. Trans. R. Soc. B: Biol. Sci.*, vol. 364, no. 1535, pp. 3549–3557 (2009 Dec.). https://doi.org/10.1098/rstb.2009.0138.

[12] 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. Virtual Real.*, vol. 1, no. 4, paper 2 (2020 Apr.). https://doi.org/10.3389/frvir.2020.00002.

[13] R. Skarbez, Jr. F. P. Brooks, and M. C. Whitton, "A Survey of Presence and Related Concepts," *ACM Comput. Surv.*, vol. 50, no. 6, paper 96 (2017 Nov.). https://doi.org/10.1145/3134301.

[14] A. Lindau and S. Weinzierl, "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acust. un. Acust.*, vol. 98, no. 5, pp. 804–810 (2012 Sep./Oct.). https://doi.org/10.3813/AAA.918562.

[15] C. Pike, F. Melchior, and T. Tew, "Assessing the Plausibility of Non-Individualised Dynamic Binaural Synthesis in a Small Room," in *Proceedings of the 55th International AES Conference* (2014 Aug.), paper 6-1.

[16] F. Brinkmann, L. Aspöck, D. Ackermann, et al., "A Round Robin on Room Acoustical Simulation and Auralization," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2746–2760 (2019 Apr.). https://doi.org/10.1121/1.5096178.

[17] D. Ackermann, F. Fiedler, F. Brinkmann, M. Schneider, and Weinzierl Stefan, "On the Acoustic Qualities of Dynamic Pseudobinaural Recordings," *J. Audio Eng. Soc.*, vol. 68, no. 6, pp. 418–427 (2020 Jun.). https://doi.org/10.17743/jaes.2020.0036.

[18] A. Neidhardt and A. M. Zerlik, "The Availability of a Hidden Real Reference Affects the Plausibility of Position-Dynamic Auditory AR," *Front. Virtual Real.*, vol. 2, no. 9, paper 678875 (2021 Sep.). https://doi.org/10.3389/frvir.2021.678875.

[19] F. Zotter and M. Frank, *Ambisonics A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* (Springer, Cham, Switzerland, 2019). https://doi.org/10.1007/978-3-030-17207-7.

[20] J. Merimaa and V. Pulkki, "Spatial Impulse Response Rendering," in *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx)*, pp. 139–144 (Naples, Italy) (2004 Oct.).

[21] L. McCormack, A. Politis, O. Scheuregger, and V. Pulkki, "Higher-Order Processing of Spatial Impulse Responses," in *Proceedings of the 23rd International Congress on Acoustics (ICA)*, pp. 4909–4916 (Aachen, Germany) (2019 Sep.).

[22] T. Lübeck, J. M. Arend, and C. Pörschmann, "Binaural Reproduction of Dummy Head and Spherical Microphone Array Data—A Perceptual Study on the Minimum Required Spatial Resolution," *J. Acoust. Soc. Am.*, vol. 151, no. 1, pp. 467–483 (2021 Jan.). https://doi.org/10.1121/10.0009277.

[23] I. Engel, C. Henry, S. V. Amengual Garí, et al., "Perceptual Comparison of Ambisonics-Based Reverberation Methods in Binaural Listening," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 121–126 (Paris, France) (2019 Sep.). https://doi.org/10.25836/sasp.2019.11.

[24] I. Engel, C. Henry, S. V. Amengual Garí, P. W. Robinson, and L. Picinali, "Perceptual Implications of Different Ambisonics-Based Methods for Binaural Reverberation," *J. Acoust. Soc. Am.*, vol. 895, no. 149, pp. 895–910 (2021 Feb.). https://doi.org/10.1121/10.0003437.

[25] M. Frank and M. Brandner, "Perceptual Evaluation of Spatial Resolution in Directivity Patterns," in *Proceedings of the 45th DAGA*, pp. 74–77 (Rostock, Germany) (2019 Mar.).

[26] A. Ahrens, K. D. Lund, M. Marschall, and T. Dau, "Sound Source Localization With Varying Amount of Visual Information in Virtual Reality," *PLOS ONE*, vol. 14, no. 3, paper e0214603 (2019 Mar.). https://doi.org/10.1371/journal.pone.0214603.

[27] W. Bailey and B. M. Fazenda, "The Effect of Visual Cues and Binaural Rendering Method on Plausibility in Virtual Environments," presented at the *144th Convention of the Audio Engineering Society* (2018 May), paper 9921.

[28] P. Larsson, D. Västfjäll, P. Olsson, and M. Kleiner, "When What You Hear is What You See: Presence and Auditory-Visual Integration in Virtual Environments," in *Proceedings of the 10th International Workshop on Presence*, pp. 11–18 (Barcelona, Spain) (2007 Oct.).

[29] P. Zahorik, "Audio/Visual Interaction in the Perception of Sound Source Distance," in *Proceedings of the 23rd International Congress on Acoustics*, pp. 7927–7931 (Aachen, Germany) (2019 Sep.). http://publications.rwth-aachen.de/record/770157.

[30] S. Riedel, M. Frank, and F. Zotter, "Localization of Real and Virtual Sound Sources in a Real Room: Effect of Auditory and Visual Cues," unpublished document (2023 Nov.). https://doi.org/10.13140/RG.2.2.14779.59680/1.

[31] T. Weber, T. Lübeck, and C. Pörschmann, "Evaluating the Influence of Different Generic Head Related Transfer-Functions on Plausibility of Binaural Rendering," in *Proceedings of the 50th DAGA*, pp. 1085–1088 (Hanover, Germany) (2024 Mar.).

[32] T. Lübeck and C. Pörschmann, "Evaluating the Plausibility of Non-Individual Head-Related Transfer Functions in Anechoic Conditions," in *Proceedings of the 10th Convention of the European Acoustics Association*, pp. 6205–6211 (Turin, Italy) (2023 Sep.).

[33] T. Carpentier, M. Noisternig, and O. Warusfel, "Twenty Years of Ircam Spat: Looking Back, Looking Forward," in *Proceedings of the 41st International Computer Music Conference (ICMC)*, pp. 270–277 (Denton, TX) (2015 Sep.).

[34] P. Majdak, Y. Iwaya, T. Carpentier, et al., "Spatially Oriented Format for Acoustics: A Data Exchange Format Representing Head-Related Transfer Functions," presented at the *134th Convention of the Audio Engineering Society* (2013 May), paper 880.

[35] K. Watanabe, P. Majdak, et al. Y. Iwaya, "SOFA: Spatially Oriented Format for Acoustics," IEICE Tech. Rep. EA2014-8, vol. 114, no. 112, pp. 19–23 (2014 Jun.).

[36] J. M. Arend, M. Ramírez, H. R. Liesefeld, and C. Pörschmann, "Do Near-Field Cues Enhance the Plausibility of Non-Individual Binaural Rendering in a Dynamic Multimodal Virtual Acoustic Scene?" *Acta Acust.*, vol. 5, no. 3, paper 55 (2021 Dec.). https://doi.org/10.1051/aacus/2021048.

[37] J. M. Arend, A. Neidhardt, and C. Pörschmann, "Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set," in *Proceedings of the 29th Tonmeistertagung - VDT International Convention*, pp. 356–363 (Cologne, Germany) (2016 Nov.).

[38] H. Stanislaw and N. Todorov, "Calculation of Signal Detection Theory Measures," *Beh. Res. Meth. Instr. Comput.*, vol. 31, no. 1, pp. 137–149 (1999 Mar.). https://doi.org/10.3758/BF03207704.

[39] R. D. Morey and J. N. Rouder, "Bayes Factor Approaches for Testing Interval Null Hypotheses," *Psychol. Meth.*, vol. 16, no. 4, pp. 406–419 (2011 Dec.). https://doi.org/10.1037/a0024377.

[40] C. Pörschmann, J. M. Arend, and R. Gillioz, "How Wearing Headgear Affects Measured Head-Related Transfer Functions," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 49–54 (Paris, France) (2019 Sep.).

[41] V. Erbes, M. Geier, H. Wierstorf, and S. Spors, "Free Database of Low-Frequency Corrected Head-Related Transfer Functions and Headphone Compensation Filter," presented at the *142nd Convention of the Audio Engineering Society* (2017 May), e-Brief 325.

[42] S. Andraszewicz, B. Scheibehenne, J. Rieskamp, et al., "An Introduction to Bayesian Hypothesis Testing for Management Research," *J. Manag.*, vol. 41, no. 2, pp. 521–543 (2015 Feb.). https://doi.org/10.1177/0149206314560412.

[43] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral Equalization in Binaural Signals Represented by Order-Truncated Spherical Harmonics," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4087–4096 (2017 Jun.). https://doi.org/10.1121/1.4983652.

[44] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in *Proceedings of the 39th DAGA*, pp. 592–595 (Merano, Italy) (2013 Mar.).

[45] A. W. Mills, "On the Minimum Audible Angle," *J. Acoust. Soc. Am.*, vol. 30, no. 4, pp. 237–246 (1958 Apr.). https://doi.org/10.1121/1.1909553.

[46] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural Rendering of Ambisonic Signals by Head-Related Impulse Response Time Alignment and a Diffuseness Constraint," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616–3627 (2018 Jun.). https://doi.org/10.1121/1.5040489.

[47] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Proceedings of the 44th DAGA*, pp. 339–342 (Munich, Germany) (2018 Mar.).

[48] M. B. Gardner, "Some Monaural and Binaural Facets of Median Plane Localization," *J. Acoust. Soc. Am.*, vol. 54, no. 6, pp. 1489–1495 (1973 Dec.). https://doi.org/10.1121/1.1914447.

[49] A. Neidhardt and N. Knoop, "Binaural Walk-Through Scenarios With Actual Self-Walking Using an HTC Vive Real-Time Rendering of Binaural Audio Interactive Binaural Audio Scenes Creating Scenes Allowing Self-Translation," in *Proceedings of the 43rd DAGA*, pp. 283–286 (Kiel, Germany) (2017 Mar.).

[50] S. V. Amengual Garí, J. M. Arend, P. Calamia, and P. W. Robinson, "Optimizations of the Spatial Decomposition Method for Binaural Reproduction," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 959–976 (2020 Dec.). https://doi.org/10.17743/jaes.2020.0063.

[51] T. McKenzie, N. Meyer-Kahlen, S. J. Schlecht, and T. Lokki, "Transfer-Plausibility of Binaural Rendering With Different Real-World References," in *Proceedings of the 48th DAGA*, pp. 154–157 (Stuttgart, Germany) (2022 Oct.).

[52] S. A. Wirler, N. Meyer-Kahlen, and S. J. Schlecht, "Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes," in *Proceedings of the 3rd AES International Conference on Audio for Virtual and Augmented Reality* (2020 Aug.), paper 3-4.

THE AUTHORS



Tim Lübeck



Tobias Weber



Christoph Pörschmann

Tim Lübeck received his B.Sc. degree in Electrical Engineering in 2017 and his M.Sc. degree in Communication Engineering in 2019 from TH Köln, Cologne, Germany. He completed his master's thesis in cooperation with the Division of Applied Acoustics at Chalmers University in Gothenburg. Since 2019, he has been a Research Fellow and working toward a Ph.D. degree at TH Köln and TU Berlin in the field of virtual acoustics, binaural technology, auditory perception, and audio signal processing.

•

Tobias Weber received his Bachelor of Science degree in Media Technology in 2022 and his Master of Science degree in Media technology in 2024 from TH Köln in Cologne, Germany. Since 2024, he has been a Research Fellow and working toward a Ph.D. at TH Köln, Germany, and TU Berlin, Germany. His research interests lie in the areas of virtual acoustic environments, audio signal processing, and auditory perception.

•

Christoph Pörschmann studied Electrical Engineering at the Ruhr-Universität Bochum (Germany) and Uppsala Universitet (Sweden). In 2001, he obtained his Doctoral Degree (Dr.-Ing.) from the Electrical Engineering and Information Technology Faculty of the Ruhr-Universität Bochum as a result of his research at the Institute of Communication Acoustics. Since 2004, he has been Professor of Acoustics at TH Köln (Germany). His research interests are in the field of virtual acoustics, spatial hearing, and the related perceptual processes.