Citation: The Journal of the Acoustical Society of America **142**, 1784 (2017); doi: 10.1121/1.5005606
View online: https://doi.org/10.1121/1.5005606
View Table of Contents: https://asa.scitation.org/toc/jas/142/4
Published by the Acoustical Society of America

On the authenticity of individual dynamic binaural synthesis

Fabian Brinkmann,[a] Alexander Lindau,[b] and Stefan Weinzierl
Audio Communication Group, Technical University of Berlin, Einsteinufer 17 c, D-10587 Berlin, Germany

(Received 7 February 2017; revised 11 September 2017; accepted 15 September 2017; published online 4 October 2017)

A simulation that is perceptually indistinguishable from the corresponding real sound field could be termed *authentic*. Using binaural technology, such a simulation would theoretically be achieved by reconstructing the sound pressure at a listener's ears. However, inevitable errors in the measurement, rendering, and reproduction introduce audible degradations, as it has been demonstrated in previous studies for anechoic environments and static binaural simulations (fixed head orientation). The current study investigated the authenticity of individual dynamic binaural simulations for three different acoustic environments (anechoic, dry, wet) using a highly sensitive listening test design. The results show that about half of the participants failed to reliably detect any differences for a speech stimulus, whereas all participants were able to do so for pulsed pink noise. Higher detection rates were observed in the anechoic condition, compared to the reverberant spaces, while the source position had no significant effect. It is concluded that the authenticity mainly depends on how comprehensive the spectral cues are provided by the audio content, and the amount of reverberation, whereas the source position plays a minor role. This is confirmed by a broad qualitative evaluation, suggesting that remaining differences mainly affect the tone color rather than the spatial, temporal or dynamical qualities. *© 2017 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).* https://doi.org/10.1121/1.5005606

[NX] Pages: 1784–1795

I. INTRODUCTION

Spatial hearing, i.e., the human ability to perceive three-dimensional sound, relies on evaluating the sound pressure signals arriving at the two ear drums, and the monaural and binaural cues imprinted on them by the outer ears, the head, and the human torso. These cues depend on the position and orientation of the sound source and the listener in interaction with the properties of the surrounding acoustical environment.[1] Binaural synthesis exploits these principles by reconstructing the pressure signals at a listener's ears, based on the measurement or the simulation of binaural impulse responses and a subsequent convolution with anechoic audio content.[2] If the electroacoustic signal chain (microphones, headphones) could be perfectly linearized, and if there were no measurement errors, this should result in an exact copy of the corresponding binaural sound events.[3] Early binaural simulations were mostly static, i.e., did not account for the listener's head orientation. It was shown, however, that head movements are important for sound source localization,[4] improve localization accuracy,[5] aid externalization,[6] and are naturally used when attending concerts, playing video games, or judging perceptual qualities such as source width and envelopment.[7] This fostered the development of dynamic binaural synthesis, where binaural impulse responses are exchanged according to the listener's position and head orientation in real-time.

The time-variant nature, however, poses additional challenges on binaural signal acquisition and processing as it requires an imperceivable round-trip system latency,[8] a perceptually transparent spatial discretization of the impulse response dataset,[9] and suitable approaches for interpolation during head movements of the listener.[10,11]

While each of these steps for signal acquisition and processing can be evaluated individually, it is not straightforward how to evaluate the entire signal chain of dynamic binaural synthesis in a comprehensive way. For this purpose, the *plausibility* and the *authenticity* of virtual acoustic environments were proposed as overall criteria for the simulated acoustical scene as well as for the quality of the systems they are generated with. While the plausibility of a simulation refers to the agreement with the listener's expectation toward a corresponding real event (agreement to an internal reference),[12] the authenticity refers to the perceptual identity with an explicitly presented real event (agreement to an external reference, Blauert, p. 373).[1] Even non-individual dynamic binaural simulations recorded with a dummy head have been shown to provide plausible simulations.[12,13] The involved participants, nevertheless, always reported audible differences, even if these did not help them to identify reality or simulation as such.

At least four empirical studies were concerned with the authenticity of binaural synthesis.[10,14–16] In all cases, the differences between reality and simulation were audible, even if the detection rates exceeded the guessing rate only slightly (depending on the audio content, listener expertise, and the experimental setup). All of these studies were conducted as static simulations, while the authenticity of *dynamic* binaural synthesis has not been assessed before. Moreover, previous studies were restricted to anechoic environments, and the results were always cumulated across participants and test conditions, neglecting the potential differences in the individual performance of participants and effects related to audio content or to the spatial configuration of source and receiver. Finally,

[a] Electronic mail: fabian.brinkmann@tu-berlin.de

[b] Current address: Max Planck Institute for Empirical Aesthetics, Grüneburgweg 14, D-60322 Frankfurt am Main, Germany.

an isolated test on authenticity provides little information about specific weaknesses of the binaural simulation, which might be valuable for technical improvements. In general, the literature has largely been focused on the evaluation of localization (e.g., Wightman and Kistler[17]), while other perceptual qualities, that might also be of high relevance in the context of virtual acoustic environments, were left unstudied.

In the current study, we thus combined tests of the authenticity of an individualized dynamic binaural synthesis in both anechoic and two reverberant environments with a comprehensive qualitative evaluation of 45 perceptual attributes. We aimed at designing the authenticity test to be as sensitive as possible, in order to produce practically meaningful results already at the level of individual participants, and to investigate the influence of room acoustical conditions, different source-receiver configurations, and the audio content.

The quality assessment of dynamic binaural synthesis with respect to authenticity as the strictest possible criterion is not only relevant to evaluate the performance and to identify potential shortcomings of binaural technology itself: It seems to become standard practice also to evaluate loudspeaker based reproduction systems by using a binaurally transcoded representation of the corresponding channels. This comprises the evaluation of mono/stereo loudspeaker setups[17,18] as well as loudspeaker arrays driven by sound field synthesis techniques such as wave field synthesis[19] or higher order ambisonics.[20] For this purpose, only an authentic simulation, including a natural interaction with the listener's head movements and a representation of the surrounding spatial environment, can provide a reliable and transparent reference for the perceptual evaluation of these techniques.

II. METHOD

Many sources of error might occur during measuring and reproducing binaural signals at the listener's ears. An inspection of errors that are relevant in the context of this study is given in Table I. It shows that each of them alone can already produce potentially audible artifacts, and has to be carefully controlled if aiming at an authentic simulation. We will thus discuss these error sources and possibilities to avoid them, before we outline the setup and methods used for perceptual testing in the following.

Hiekkanen *et al.*[18] reported that head movements of 1 cm to the side and azimuthal head-above-torso rotations of 2.5° already produce audible differences in binaural transfer functions. To account for this, previous studies used some kind of head rest to restrict the participants' head position, and monitored the participants' head position with optic or magnetic tracking systems.[14,15] Throughout the study, Masiero[15] allowed for head movements between ±1 cm translation, and ±2° rotation, respectively. In the current study, we allowed tolerances of ±1 cm, and ±0.5° during the measurement of the binaural transfer functions.

Similar errors are induced by repositioning the microphones,[21] or headphones[22] which was shown to be audible even for naive listeners in the case of headphone repositioning.[23] In the context of this study such problems can be avoided, if the microphones are kept in position while measuring the binaural transfer functions of loudspeakers and headphones, and if the headphones are worn during the entire experiment.

The presence of headphones, however, influences the sound field at the listener's ears because they act as an obstacle to sound arriving from the outside, as well as sound being reflected from the listener's head. This causes distortions in the magnitude and phase spectra,[10,24,25] as well as changes in the acoustic load seen from inside the ear canal (free air equivalent coupling criterium[22]). To avoid this, Langendijk and Bronkhorst[10] used small extra-aural earphones with a limited band width, while Moore *et al.*[14] had cross-talk cancelled transaural loudspeakers for binaural signal reproduction. We used extra-aural headphones with full band width, whose influence on external sound fields are comparable to the earphones used by Langendijk and Bronkhorst.[25]

Moreover, headphone transfer functions (HpTFs) show considerable distortions that need to be compensated by means of inverse filters, which are typically designed by frequency dependent regulated inversion of the HpTFs.[26] During the filter design, it is vital to find a good balance between an exact inversion that would result in filters with undesired high gains at the frequencies of notches in the HpTFs (possibly causing audible ringing artifacts), and too much regulation which is likely to cause high frequency damping.[21] To assure this, we applied regularization only at frequencies where notches in the HpTFs occurred.

Note that in the case of authenticity, the room and the loudspeakers are considered to be a part of the experiment. This is in contrast to HRTF measurements, where their influence should be removed from the measured transfer functions by means of post-processing, and thus become additional sources of error.[27]

A. Experimental setup

The listening tests were conducted in the anechoic chamber and the recording studio of the State Institute for

TABLE I. Sources of errors and variance in the measurement and reproduction of binaural signals. Typical, and maximum errors were either directly taken from the references, or obtained by visual inspection of corresponding figures.

Source	Amount typ. (max)	Reference
Head repositioning	4 (10) dB	Riederer (Ref. 44); Hiekkanen *et al.* (Ref. 18)
Microphone repositioning	5 (20) dB	Lindau and Brinkmann (Ref. 21)
Headphone repositioning	5 (20) dB	Møller *et al.* (Ref. 22); Paquier and Koehl (Ref. 23)
Acoustic headphone load	4 (10) dB	Møller *et al.* (Ref. 22)
Headphone presence	10 (25) dB	Langendijk and Bronkhorst (Ref. 10); Moore *et al.* (Ref. 24); Brinkmann *et al.* (Ref. 25)
Headphone compensation	1 (10) dB	Lindau and Brinkmann (Ref. 21)

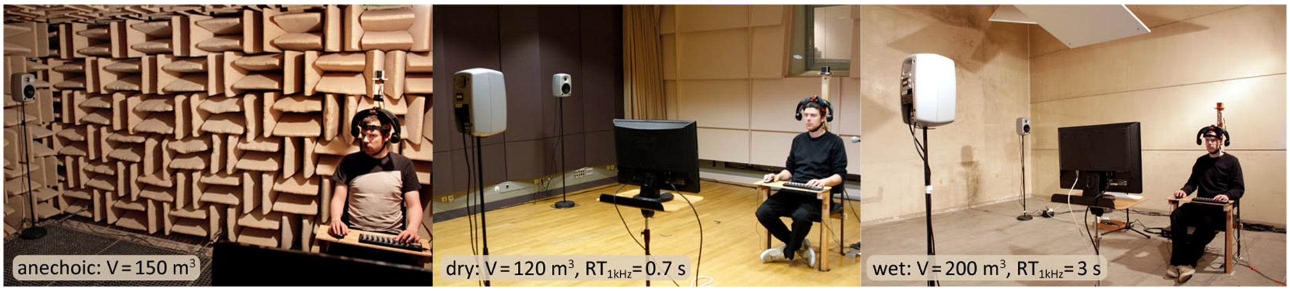
anechoic: V = 150 m³
dry: V = 120 m³, RT1kHz = 0.7 s
wet: V = 200 m³, RT1kHz = 3 s

FIG. 1. (Color online) Listening test setup in the anechoic, dry, and wet test environment.

Music Research, and in the reverberation chamber of TU Berlin (Fig. 1). The three rooms are of comparable volume but exhibit large differences in reverberation time. To limit the duration of the experiment to a practical amount, the reverberation time of the wet room was reduced from 6.7 to 3 s at 1 kHz using 1.44 m^3 porous absorber. Participants were seated on a chair equipped with a height and depth adjustable neck rest, and a small table providing an arm-rest and space for placing the MIDI interface used throughout the test (Korg nanoKONTROL). An LCD screen was used as visual interface and placed 2 m in front of the participants at eye level.

Two active near-field monitors (Genelec 8030a) were placed in front and to the right of the participants at a distance of 3 m and a height of 1.56 m, corresponding to source positions of 0° and 90° azimuth, and 8° elevation. The height was adjusted so that the direct sound path was not blocked by the LCD screen. The source positions were chosen to represent the most relevant use case of a frontal source, as well as the potentially critical case of a lateral source, where the signal to noise ratio decreases due to shadowing of the head. With a loudspeaker directivity index of ca. 5 dB at 1 kHz[28] and corresponding critical distances of 1.3 m (dry) and 0.8 m (wet), the source positions result in slightly emphasized diffuse field components in the reverberant environments.

For reproducing the binaural signals, low-noise DSP-driven amplifiers and extra-aural headphones were used, which were designed to exhibit minimal influence on sound fields arriving from external sources while providing full audio bandwidth (BKsystem[29]). To allow for an instantaneous switching between the binaural simulation and the corresponding real sound field, the headphones were worn during the entire listening test, i.e., also during the binaural measurements. The participants' head positions were controlled using head tracking with 6 degrees of freedom (x, y, z, azimuth, elevation, lateral flexion) with a precision of 0.001 cm and 0.003°, respectively (Polhemus Patriot). A long term test of 8 h showed no noticeable drift of the tracking system.

Individual binaural transfer functions were measured at the blocked ear canal using Knowles FG-23329 miniature electret condenser microphones flush cast into conical silicone ear-molds. The molds were available in three different sizes, providing a good fit and reliable positioning for a wide range of individuals.[21] Phase differences between left and right ear microphones did not exceed ±2° below 1 kHz to avoid audible interaural phase distortion.[30]

The experiment was monitored from a separate room with talk-back connection to the test environment.

B. Individual transfer function measurement

Binaural room impulse responses (BRIRs) and HpTFs were measured and processed for every participant prior to the listening test. MATLAB and AKtools[31] were used for signal generation, playback, recording, and processing at a sampling rate of 44.1 kHz. The head positions of the participants were displayed using Pure Data. Communication between the programs was done by UDP messages.

Before starting, participants put on the headphones and were familiarized with the measurement procedure. Their current head position, given by azimuth and x/y/z coordinates was displayed on the LCD screen along with the target azimuth. The head tracker was calibrated with the participant looking at a frontal reference position marked on the LCD screen. Participants were instructed to keep their eye level aligned to the reference position during measurement and listening test, this way establishing also an indirect control over their head elevation and roll. For training proper head-positioning, participants were instructed to move their head to a specific azimuth and hold the position for 10 s. A visual inspection showed that all participants were quickly able to maintain a position with a precision of ±0.2° azimuth, and ±2 mm translation in x/y/z coordinates.

Then, participants inserted the ear-molds with measurement microphones into their ear canals until they were flush with the bottom of the concha, and the correct fit was inspected by the investigator. BRIRs were measured for azimuthal head-above-torso orientations within ±34° in 2° steps providing a perceptually smooth adaption to head movements.[9] The range allowed for a convenient view of the LCD screen at any head orientation. Sine sweeps of an FFT order 18 were used for measuring transfer functions, with the level of the measurement signal being identical across participants. It was set so to avoid limiting of the DSP-driven loudspeakers, and headphones, and to achieve a peak-to-tail signal-to-noise ratio (SNR) of approximately 80 dB for ipsilateral and 60 dB for contralateral sources without averaging.[32] Because the ear-molds significantly reduced the level at the ear drums, all participants reported it to be still comfortable.

The participants started the measurement by pressing a button on the MIDI interface after moving their head to the target azimuth with a precision of ±0.1°. For the frontal head orientation, the reference position had to be met within 0.1 cm for the x/y/z-coordinates. For all other head orientations the translational positions naturally deviate from zero; in these cases, participants were instructed to meet the targeted azimuth only, and to move their head in a natural way. During the measurement, head movements of more than 0.5° or 1 cm caused a repetition, which rarely happened. The tolerances were set to avoid audible artifacts introduced by imperfect positioning.[1,18]

Thereafter, ten individual HpTFs were measured per participant. Although the headphones were worn during the entire experiment, their position on the participants' head might change due to head movements. To account for corresponding changes in the HpTFs, participants were instructed to rotate their head to the left and right in between measurements. After the measurements, which took about 30 min, the investigator carefully removed the in-ear microphones without changing the position of the headphones.

C. Postprocessing

As a first step, leading zeros in the BRIRs were removed, while the temporal structure remained unchanged. For this purpose, time-of-arrivals (TOAs) were estimated using onset detection, and removed by means of a circular shift. TOA outliers were corrected by fitting a second order polynom or smoothing splines to the TOA estimates—whatever gave the best fit to the valid data (determined by visual inspection). ITDs, i.e., differences between left and right ear TOAs, were re-inserted in real time during the listening test to avoid comb-filter effects occurring in dynamic auralizations with non-time-aligned BRIRs and reducing the overall system latency.[11] In a second step, BRIRs were truncated to 0.4, 1, and 3 s for the anechoic, dry and wet environment to allow for a decay of around 60 dB. A squared sine fade out was applied at the intersection between the impulse response decay and the noise floor to artificially extend the decay.

Individual HpTF compensation filters of FFT order 12 were designed based on the average HpTF using frequency dependent regularized least mean squares inversion.[26] Regularization was used to limit filter gains if perceptually required: HpTFs typically show distinct notches at high frequencies which are most likely caused by anti-resonances of the pinna cavities.[33] For an example see Fig. 2 (top) at approximately 10 and 16 kHz. The exact frequency and depth of these notches strongly depends on the current fit of the headphones. Already a slight change in position might considerably detune a notch, potentially leading to ringing artifacts of the applied headphone filters.[21] Therefore, individual regularization functions were composed by manually fitting one to three parametric equalizers (PEQs) per ear to the most disturbing notches. The compensated headphones approached a minimum phase target band-pass consisting of a 4th order Butterworth high-pass with a cut-off frequency of 59 Hz and a second order Butterworth low-pass with a cut-off frequency of 16.4 kHz. The result, i.e., the convolution of each HpTF with the inverse filter, deviated from the target band-pass by less than ±0.5 dB in almost all cases, except for frequencies where notches in the HpTF occurred (cf. Fig. 2). The frequency responses of the in-ear microphones remained uncompensated in the BRIRs and HpTFs. This way the inverse frequency responses are present in the HpTF filters, and the microphones influence cancels out if the HpTF filters are convolved with the BRIRs.

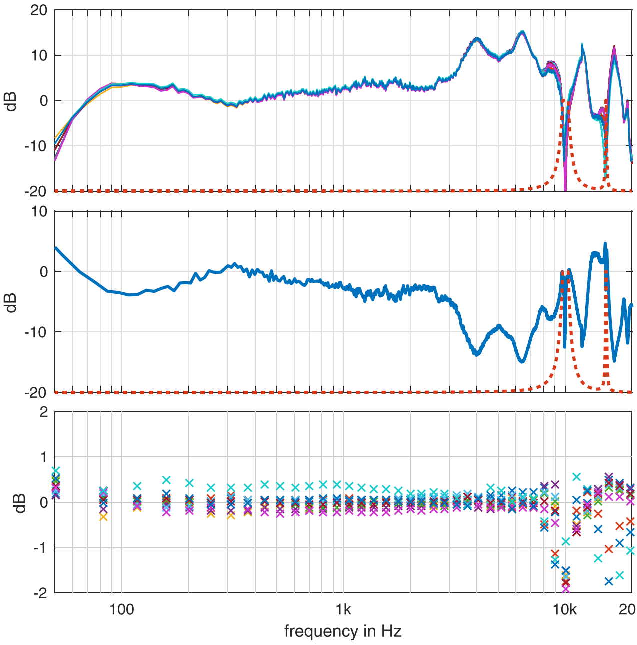

FIG. 2. (Color online) Example of the headphone compensation process for the left ear of participant 3. (Top) HpTFs (solid lines) and regularization (dashed line). (Middle) Compensation filter (solid line) and regularization (dashed line). (Bottom) Difference between compensated HpTFs and target band pass in auditory filters.

Finally, presentations of the real loudspeaker and the binaural simulation had to be matched in loudness. Assuming that signals obtained via individual binaural synthesis closely resemble those obtained from loudspeaker reproduction in the temporal and spectral shape (cf. Fig. 3), loudness matching can be achieved by simply matching the RMS-level of simulation and real sound field. Hence, 5 s pink noise samples were recorded from loudspeakers and headphones while the participant's head was in the frontal reference position. Before matching the RMS-level, the headphone recordings were convolved with the frontal incidence BRIR and the headphone compensation filter to account for the reproduction paths during the listening test. The loudspeaker recordings were convolved with the target bandpass that was used for designing the headphone compensation filter.

D. Test procedure

Nine participants with an average age of 30 years (6 male, 3 female) participated in the listening test, all of them experienced with dynamic binaural synthesis. No hearing anomalies were known, and with a musical background of 13 years on average, all participants were regarded as expert listeners.

The test procedure was identical across the three acoustical environments, and tests were conducted over a period of 20 months. At first, participants were placed on the chair,

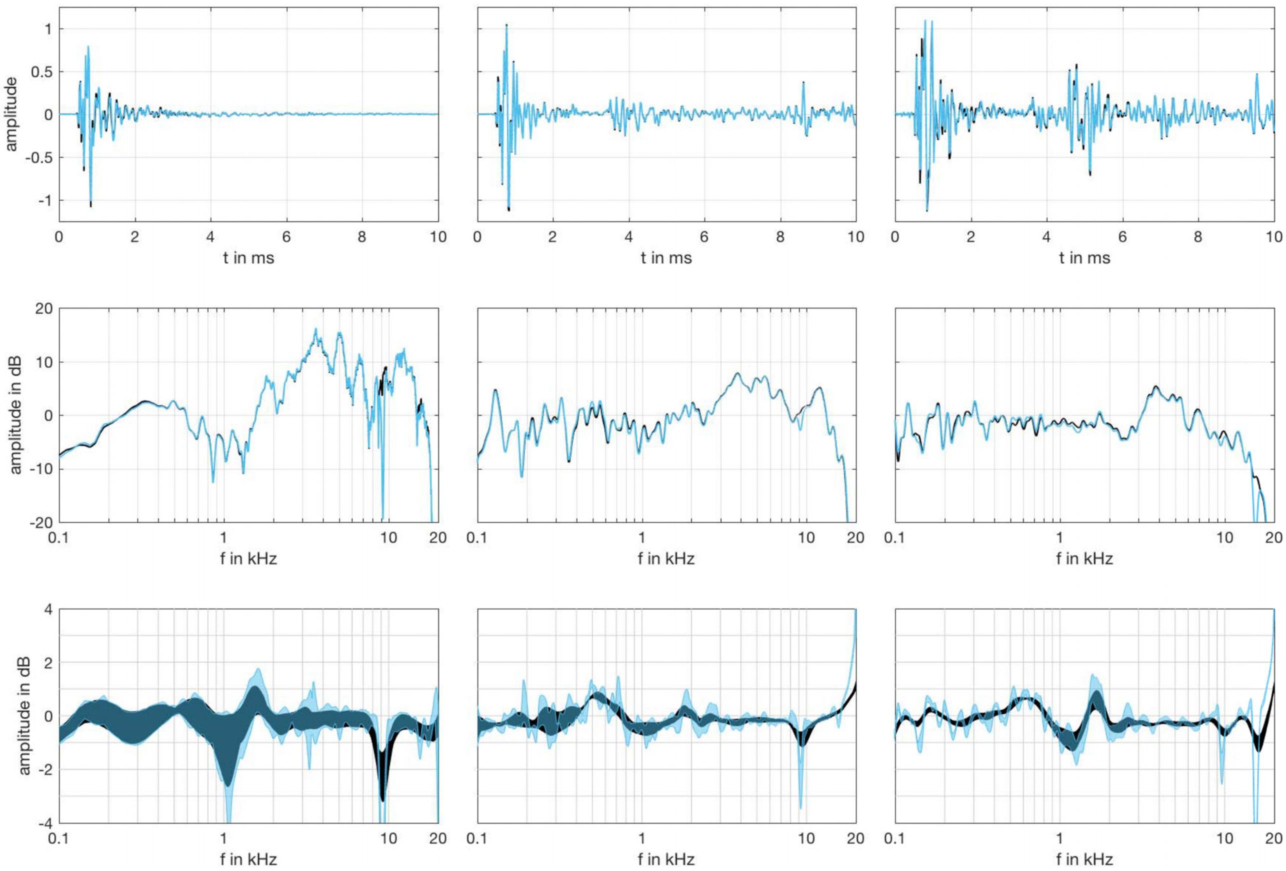

FIG. 3. (Color online) Differences between binaural simulation and real sound field for the frontal source measured in the anechoic (left), dry (middle), and wet (right) acoustic environment. Top row shows real (black lines) and simulated (blue lines) binaural impulse responses, middle row real and simulated binaural magnitude spectra for a neutral head-above-torso orientation (12th octave smoothed spectra are shown in case of the dry and wet environment to improve readability). Bottom row shows the range of differences between 12th (light blue) and 3rd octave (dark blue) smoothed magnitude spectra for all head-above-torso orientations.

took on the headphones, and were familiarized with the user interfaces showing the current head position and answer buttons, and the MIDI interface for their control. Afterward, the BRIRs and HpTFs were measured and processed as described above.

The perceptual testing started with four ABX tests for authenticity per participant (2 sources × 2 contents) each consisting of 24 trials. The order of content and source was randomized and balanced across participants. At each trial, the binaural simulation and the real sound field were randomly assigned to three buttons (A/B/X, with each condition assigned at least once), and participants started and stopped the audio playback by pressing a button on the MIDI interface. Stopping the playback could also be used to listen to the entire decay in the BRIR. The ABX test is a 3-Interval/2-Alternative Forced Choice (3I/2AFC) test, with the three intervals *A*, *B*, and *X*, and the two possible answers (forced choices) *A equals X*, and *B equals X*.

The participants could take their time at will to repeatedly listen to *A*, *B*, and *X* in any order and switch at any time. They were moreover instructed to listen at different azimuthal head-above-torso orientations, to focus on different frequency ranges, and that dynamic cues induced by head movements might also help to distinguish between simulation and reality. Because it was not clear which kind of head movements or positions would be helpful, it was left to the participants to find the best head positions/movements for detecting differences. To avoid a drift in the positioning of the participants during the experiment, they were instructed to keep their head at approximately 0° elevation throughout the test, and to move their head to the reference position given by azimuth and x/y/z coordinates between trials. To ensure this, the participants' head position was monitored by the experimenter, who manually enabled each trial. In addition, head positions were recorded in intervals of 100 ms for *post hoc* inspection (cf. Sec. III B).

Pulsed pink noise and an anechoic male speech recording (5 s) were used as audio content. Speech was chosen as a familiar "real-life" stimulus including transient components that were supposed to reveal potential flaws in the temporal structure of the simulation. Noise pulses were believed to best reveal flaws related to the spectral shape. To allow for establishing a stable impression of coloration and decay, a single noise pulse with a length of 0.75 s followed by 1 s silence (anechoic and dry environment), and a length of 1.5 s followed by 2 s silence (wet environment) was played in a loop. Noise bursts were faded in and out with a 20 ms squared sine window. The bandwidth of the stimuli was restricted using a 100 Hz high-pass to eliminate the influence of low frequency background noise on the binaural transfer functions. Previous studies (cf. studies C and D in Table III) obtained almost identical detection rates for speech and music, which was confirmed by informal listening prior to the current study. We thus limited ourselves to two types of audio content, in order to allow for more variation in other independent variables (source position, spatial environment).

In a next step, qualitative differences between binaural simulation and real sound field were assessed using the Spatial Audio Quality Inventory (SAQI[34]) as implemented in the WhisPER toolbox.[35] Again, participants could directly compare the two test conditions and take their time at will before giving an answer. Audio playback was started and stopped using two buttons labeled A and B, behind which the simulation and real sound field were hidden, i.e., participants did not know which button toggled the real sound field. The presentation order of the qualities was randomized to avoid order effects. A list with the names, and descriptions of the perceptual qualities was given to all participants beforehand, and questions could be discussed on site. In addition, attributes and their description were also displayed on the screen to avoid any misunderstandings.

The test took about 2 h including breaks during which the participants had to remain seated to avoid any change in the test environment that might have introduced additional errors; 30 min were needed for binaural measurements, 10 min for post processing. The participants took on average 50 min for AFC testing ($SD = 11$ min), and 21 min for the SAQI ratings ($SD = 9$ min). The test duration was perceived as just about tolerable by the participants.

Dynamic auralization was realized using the fast convolution engine *fWonder*[36] in conjunction with an algorithm for real-time reinsertion of the ITD.[11] *fWonder* was also used for applying (a) the HpTF compensation filter and (b) the loudspeaker target bandpass. The playback level for the listening test was set to 60 dB sound pressure level $(\text{SPL})_{\text{AFeq}}$. This way the 60 dB dynamic range in the measured BRIRs ensures that their decay continues approximately until the absolute threshold of perception around 0 dB SPL is reached. Any artifacts related to the truncation of the measured BRIRs were thus expected to be inaudible. BRIRs used in the convolution process were dynamically exchanged according to the participants' current azimuthal head-above-torso orientation (head azimuth), and playback was automatically muted if the participant's head orientation exceeded 35° azimuth.

E. Alternative forced choice test design

The M-I/N-AFC method provides an objective, criterion-free, and particularly sensitive test for the detection of small differences,[37] and thus seems appropriate for testing the authenticity of virtual environments. As a Bernoulli experiment with a guessing rate of 1/N, the binomial distribution allows to calculate the probability that a certain number of correct answers occurs by chance, thus enabling tests on statistical significance: If the amount of correct answers is significantly above chance level, the simulation would *not* be considered as perceptually authentic.

If N-AFC tests are used in the context of authenticity, one should be aware that this corresponds to proving the null hypothesis H_0, i.e., proving that simulation and reality are indistinguishable. Strictly speaking, this proof cannot be given by inferential statistics. The approach commonly pursued is to establish empirical evidence that supports the H_0 by rejecting a minimum effect alternative hypothesis H_1 representing an effect of irrelevant size, i.e., a tolerable increase of the detection rate above the guessing rate.[38]

The test procedure is usually designed to achieve small type 1 error levels (wrongly concluding that there was an audible difference although there was none) of typically 0.05, making it difficult—especially for smaller differences—to produce significant test results. If we aim, however, at proving the H_0 such a design may unfairly favor our implicit interest ("progressive testing"), that is reflected in the type 2 error (wrongly concluding that there was no audible difference although indeed there was one).

Therefore, we first specified a practically meaningful detection rate of $p_{\text{d}} = 0.9$ to be rejected, and then aimed at balancing type 1 and type 2 error levels in order to statistically substantiate the rejection *and* the acceptance of the null hypothesis, i.e., the conclusion of authenticity. According to these considerations, a 3I/2AFC listening test design with 24 trials for each participant and test condition was chosen. This lead to a critical value of 18 (75%) or more correct answers in order to reject the $H_0(p_{2\text{AFC}} = 0.5)$, while for less than 18 correct answers, the specific $H_1(p_{2\text{AFC}} = 0.9)$ could be rejected (p_{NAFC}: N-AFC detection rate). Type 1 and type 2 error levels were initially set to 5% and corrected for multiple testing of 4 test conditions by means of Bonferroni correction.

The detection rate of $p_{2\text{AFC}} = 0.9$ may seem high at first glance, but it corresponds to the expectation that even small differences would lead to high detection rates, considering trained participants and a sensitive test procedure (cf. Leventhal,[37] p. 447) that included suitable audio contents and unlimited listening. Moreover, the critical value of 18 corresponds to the threshold of perception where a participant would identify existing differences in 50% of the cases, which seems to be an adequate criterion for deciding whether or not a simulation is authentic. Note that N-AFC detection rates can be corrected for guessing by $[p_{\text{NAFC}} - 1/N] \cdot [1/(1 - 1/N)]$.

F. Qualitative test design

The German version of the Spatial Audio Quality Inventory (SAQI) was used for assessing detailed qualitative judgements. It consists of 48 perceptual attributes for the evaluation of virtual acoustic environments which were elicited in an expert focus group for virtual acoustic environments. Each SAQI quality is accompanied by a short verbal description as well as suitable scale end labels.

We used the SAQI for a direct comparison, i.e., participants rated differences between the simulation and the real sound field, with a rating of zero indicating no perceivable difference. As we were interested in a broad and explorative evaluation, only three qualities of the complete SAQI were excluded, because they were considered irrelevant in our case (*Speed*, *Sequence of events*, *Speech intelligibility*). To limit the time of the listening test to a practical amount, the qualitative evaluation was only carried out for the frontal sound source and the pulsed pink noise.

III. RESULTS

A. Physical evaluation

Prior to the perceptual evaluation, acoustic differences between the test conditions were estimated based on measurements with the FABIAN dummy head that is equipped with a computer controllable neck joint.[36] Therefore, FABIAN was placed on the chair to measure BRIRs and HpTFs as described in Sec. II. In a second step, BRIRs were measured as being reproduced by the headphones and the simulation engine as described above. Differences between simulation and real sound field for the left ear and the frontal source are shown in Fig. 3. They are comparable to the right ear differences, and the lateral source.

Simulated and real BRIRs for the neutral head-above-torso orientation (top row), show a striking similarity for all test environments. For ease of display only the first 10 ms are shown. Corresponding magnitude spectra (middle row) are very similar for the anechoic environment. Slightly higher deviations occur for the reverberant environments in certain frequency ranges (e.g., around 1 kHz), presumably caused by differences in the late part of the BRIRs.

For a better overview, the range of errors for all head-above-torso orientations between ±34° is illustrated in the bottom row of Fig. 3 for 3rd and 12th octave smoothed magnitude spectra. For most frequencies and head orientations, differences are in the range of ±1 dB which is in good accordance to results of earlier studies.[3,10,14,39] Larger deviations occur at frequencies of about 9 kHz and 16 kHz where narrow and deep notches in the HpTF remained uncompensated for robustness against headphone re-positioning (cf. Sec. II C). However, they exhibit widths of 9% or less relative to their center frequency, and were thus expected to be inaudible; Moore *et al.*[40] reported an audibility threshold of 12.5% relative notch width. Spectral differences are slightly larger in the anechoic environment, in particular at about 1 kHz. At this frequency a notch appears for the left ear in case the head is turned away from the source—i.e., for head-above-torso orientations in the range of 30°. This notch originates from delayed copies of the sound traveling around the head on different paths.[41]

Assuming that third octave differences in the range of 0.5 dB might already be audible for expert listeners and sensitive listening test designs (compare ΔG_{95} in Table III from Ref. 42), we can expect that the binaural simulation will turn out to be not perceptually authentic, at least for the noise content.

B. Perceptual authenticity

The detection rates of the 2AFC test are summarized in Fig. 4 for all participants and test conditions. Although statistical analysis of authenticity was conducted on the level of individual participants, the observed average detection rates are given in Table II(a) for better comparability to earlier studies, and because the corresponding detection frequencies were used to statistically analyze effects between test conditions by means of χ^2 tests. One participant could not participate in the anechoic environment due to illness, and two participants who accidentally touched the headphones after the binaural measurements were excluded from the results of the dry environment. Both reported hissing sounds that might be attributed to ringing artifacts caused by headphone repositioning.

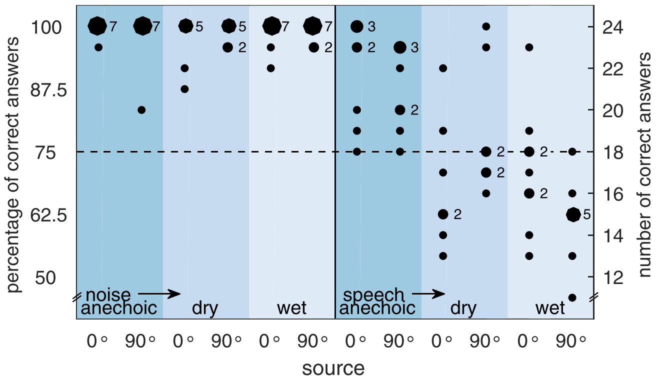

FIG. 4. (Color online) Detection rates of the 2AFC test for all participants and test conditions. The size of the dots and the numbers next to them indicates how many participants scored identical results. Results on or above the dashed line are significantly above chance, indicating that differences between simulated and real sound field were audible.

A clear difference in detection performance was found between the audio contents: for pulsed noise, all participants were able to discriminate simulation and real sound field, i.e., all individual detection rates are above the dashed line in Fig. 4. For the speech stimulus, however, the simulation turned out to be authentic in 44% of cases (dots below dashed line). χ^2 tests showed this effect to be statistically highly significant ($\chi^2 = 32.81$, $p < 0.001$, $df = 1$). A significant effect of the room was observed in interaction with the speech content, where all participants detected differences in the anechoic environment, whereas only 43%, and 28% detected differences in the dry and the wet environment, respectively ($\chi^2 = 8.46$, $p = 0.001$, $df = 2$). Pairwise comparisons showed significant differences between the anechoic and wet room ($\chi^2 = 7.96$, $p = 0.005$, $df = 1$) and almost

TABLE II. Data from the 2AFC listening test averaged across participants. (a) 2AFC detection rates in percent. (b) Rating duration per trial in seconds (additionally averaged across trials). (c) Amount of head movements specified by the difference P_{75}–P_{25} of observed head azimuths in degree (P_i: *i*th percentile).

		Anechoic		Dry		Wet		
		0°	90°	0°	90°	0°	90°	All
II(a)								
	Noise	99.5	97.9	97.0	98.8	98.6	99.1	98.5
	Speech	91.2	87.5	68.5	79.1	71.3	61.6	76.2
II(b)								
	Noise	10.3	9.1	15.2	15.1	12.3	13.2	12.4
	Speech	30.3	29.6	43.4	39.1	35.9	40.7	36.3
II(c)								
	Noise	6.6	5.6	8.5	9.1	10.6	12.6	8.9
	Speech	21.6	16.6	18.6	28.2	27.3	19.0	21.8

significant differences between the anechoic and dry room ($\chi^2 = 3.57$, $p = 0.059$, $df = 1$). Differences between the dry and wet room were statistically insignificant with the given sample size and test power ($\chi^2 = 0.88$, $p = 0.349$, $df = 1$). In line with earlier studies, no significant effect was found for the source position ($\chi^2 = 0.04$, $p = 0.84$, $df = 1$).

Differences between audio contents were also found in the rating durations and the amount of azimuthal head movements. An inspection reveals that giving an answer for a trial took the participants about three times longer when listening to the speech content [Table II(b)], and that they used larger head movements to detect differences [Table II(c)]. Both effects are highly significant in all environments (Bonferroni corrected Wilcoxon signed rank tests for dependent samples, $p \leq 0.01$, $\alpha = 0.05$). This is also reflected by the high and significant correlations between the average detection rates [Table II(a)] and (1) the rating duration (Pearson correlation, $r = -0.92$, $p < 0.01$), and (2) the amount of head movements ($r = -0.73$, $p < 0.01$): Participants with lower detection rates took more time and moved their head further when trying to detect differences between reality and simulation. Participants who could not reliably detect differences, on average explored 95% of the available range of head-above-torso orientations. Thus, it is unlikely that the results are biased due to an insufficient exploration of the binaural simulation. Again no effects for the source position were observed, whereas effects of the acoustic environment are rather small but significant at least for the rating duration between the anechoic and dry, as well as the anechoic and wet room (Bonferroni corrected Wilcoxon signed rank tests for dependent samples, $p \leq 0.05$, $\alpha = 0.05$).

During auralization, BRIRs were selected solely by the participants' head azimuth. Hence, unobserved differences with respect to the measured positions in the remaining degrees of freedom—translation in x, y, z, elevation, lateral flexion—might have caused audible artifacts. Therefore, the recorded head positions of all participants were used for a *post hoc* analysis of deviations between head position during binaural measurements and 2AFC tests. For the translation in x, y, and z coordinates, deviations were found to be smaller than 1 cm for about 95% of the time and never exceeded 2 cm. Differences in head elevation (tilt) and in lateral flection (roll) rarely exceeded 10° and were below 5° for 90% of the time. While this may have caused audible artifacts occasionally, a systematic influence of the results is highly unlikely (cf. Ref. 1, p. 44 and Ref. 18).

C. Qualitative evaluation

The qualitative analysis with respect to 45 perceptual attributes is summarized in Fig. 5(a). They were only assessed for the frontal source and the pulsed pink noise to limit the duration of the listening test. Please note that the scale labels (y labels) were omitted for better readability. They can be found in Table I in Lindau *et al.*,[34] whereby a rating of -1 refers to the first, and a rating of 1 to the second label. A rating of 0 indicates no perceptual difference between simulation and real sound field. Because the ratings were not normally distributed in 60% of the cases (Shapiro-Wilk tests, $p \leq 0.2$), Fig. 5 shows the median values, interquartile ranges (IQR), and the total range.

In line with the results of the AFC test, several findings indicate that the simulation has a high degree of realism with respect to almost all tested perceptual aspects: (1) the IQR does include zero in almost all cases; (2) the median is zero in 92% of the cases; and (3) all participants made zero ratings for seven qualities (*roughness*, *doppler effect*, *front/back position*, *pre-echos*, *noise-like artifacts*, *alien source*, *distortion*), while 14 more qualities obtained zero ratings from all participants in at least one environment (*metallic tone color*, *level of reverberation*, *duration of reverberation*, *envelopment*, *spatial disintegration*, *post-echos*, *temporal disintegration*, *responsiveness*, *dynamic range*, *compression*, *pitched artifact*, *impulsive artifact*, *ghost source*, *tactile vibration*).

Larger differences (IQRs that do not overlap zero) were found for the *difference*, which confirms the results of the 2AFC test, the *tone color bright/dark*, where the negative ratings indicate that the simulation was perceived to be darker than the real sound field in the anechoic environment, and the *horizontal direction* in the dry room. Apart from the latter two cases, IQRs overlap each other for all test conditions, suggesting that differences between rooms are rather small.

In tendency, the participants' ratings indicate that the simulation has slightly less *naturalness*, *clarity*, and *presence*, and that the real sound field was preferred over the simulation (*liking*). However, median values were zero for the above mentioned qualities, except for *liking* in the anechoic environment. Apart from that, non-zero median values, or large IQRs were only found in the categories *tone color*, *tonalness*, and *geometry*, in turn suggesting that there are no relevant deficits regarding *room*, *time*, *dynamics*, or *artifacts* of any kind.

In some cases, equally distributed positive and negative ratings could conceal perceptually relevant differences if they result in a zero median. To uncover this effect, distributions for all absolute ratings with median values ≥ 0.05 are shown in Fig. 5(b)—sorted in descending order to emphasize their relevance. Besides the overall *difference*, three perceptual qualities related to coloration (*high frequency tone color*, *tone color bright-dark*, and *pitch*), as well as *distance* show systematic deviations from zero. However, the IQRs already include zero for *pitch* and *distance*.

IV. DISCUSSION

At least four empirical studies were concerned with the authenticity of binaural simulations, i.e., with the physical and perceptual identity of ear signals produced by natural acoustic environments and their equivalent produced by binaural synthesis: Langendijk and Bronkhorst[10] (termed *A* in the following), Moore *et al.*[14] (*B*), Masiero[15] (*C*), and Oberem *et al.*[16] (*D*). In contrast to all previous studies, which used static synthesis in anechoic conditions, the current investigation considered, for the first time, dynamic binaural synthesis, allowing for natural head movements of the listeners, as well as a sample of three different acoustic

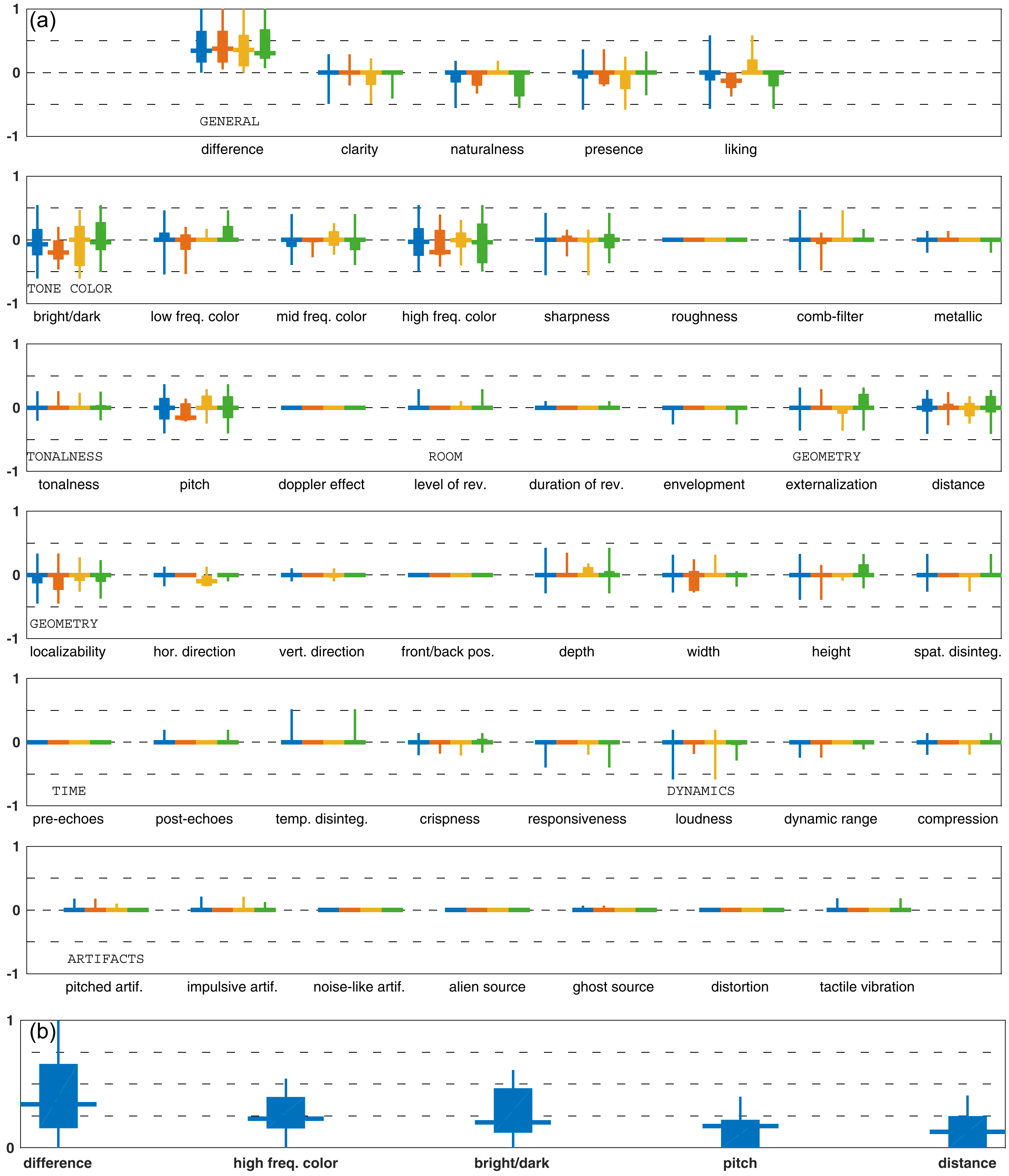

FIG. 5. (Color online) SAQI ratings by means of median (horizontal lines), IQR (boxes), and overall range (vertical lines). (a) The four bars for each perceptual quality show results pooled across rooms, and for the anechoic, dry and wet room (from left to right). (b) Pooled absolute ratings with median values ≥0.05 in descending order.

environments with different degrees of reverberation. The results can thus be expected to have more ecological validity with respect to the large variety of current and future applications of binaural technology.

The physical identity, i.e., the extent to which inaccuracies of the binaural reconstruction could be controlled, was similar in all studies. In terms of magnitude deviations between real and simulated binaural transfer functions, comparable values have been reported: *A* found 12th octave magnitude differences of ±1 dB for ipsilateral sources and ±5 dB for contralateral sources, along with phase differences of up to 6°. *B* reported magnitude deviations in the unsmoothed spectra to be smaller than ±2 dB except for frequencies above 6 kHz where deep notches occurred. We found 12th octave magnitude differences to be smaller than ±1 dB for most frequencies (cf. Fig. 3). Deviations of comparable magnitude were also observed in studies that focused *only* on the physical accuracy in the reproduction of binaural signals.[3,39,43] We can thus conclude that this seems to be the degree to which the physical identity of reality and simulation can be brought by carefully controlling the measurement, processing, and reproduction of binaural simulations.

The perceptual identity of acoustic environments and their binaural simulations could not be observed, neither in previous nor in the current study. To investigate this, all studies used M-Interval, N-Alternative Forced Choice tests (M-I/N-AFC). An overview of the results is shown in Table III, with results from *D* given as a footnote, because the test was conducted in the same laboratory as *C* with identical audio content, test environment, and test method. Since studies *C* and *D* did not conduct significance tests, the critical detection rates were computed based on the reported test method, and transformed to 2AFC rates afterward. The overview shows that (1) the detection rates were significantly above chance level except for a synthetic and strongly band limited complex tone in *B*, (2) the equivalent 2AFC detection rates for noise span from 52% in *A* (2% above chance level) to 87.5% in *B*, and (3) detection rates of previous studies were generally lower than those observed in the current study.

Since the physical accuracy was comparable in all studies, there are three factors which can account for the considerable differences in the measured detection rates. This is, (1) the presented audio content, (2) the technical implementation of the listening test, and (3) the exact test procedure of listening and decision making.

With respect to the audio content, stimuli with a broadband, steady, and non-sparse spectrum (such as noise or pulses) produced considerably and significantly higher detection rates than music or speech. This was shown by *C*, with 87.5% for noise compared to 71.4% for speech and 73.7% for music, as well as by the current study, with 98.5% for noise and 76.2% for speech. This can be attributed to a surplus of physical cues facilitating the identification of small timbral differences, which were shown to be mainly responsible for the detection performance by the qualitative evaluation. Obviously, these spectral cues were outweighed neither by the more transient character of speech, nor by the higher familiarity with speech as an everyday stimulus, since the listener does not have to draw on his or her internal reference and experience in an N-AFC listening test providing an immediate comparison between simulation and reality.

Comparing the technical setup of studies *A*, *B*, *C*, and *D*, two differences become apparent. First, the circumaural headphones used by *C*, and *D* were the only ones that had to be repositioned after measuring the binaural signals. Since even naive listeners are able to reliably detect differences due to headphone repositioning,[23] it is almost certain that this effect considerably increased the resulting detection rates. In addition, participants in *D* were re-seated between binaural measurements and listening test, which is also likely to introduce audible artifacts.[18] To avoid this, we used extra-aural headphones that remained in position during the entire listening test, and allowed for removing the in-ear microphones without moving the headphones. Moreover, participants remained in position during the entire listening test, and their position was monitored with a high precision head tracking system.

With respect to the test method, participants could listen only once to a sequence of stimuli before giving an answer

TABLE III. Overview of studies on perceptual authenticity of binaural synthesis. Sorted by ascending detection rates from left to right, and named according to the first author. For the current study, only the results for the anechoic environment are listed. See text for details.

	Langendijk (A)	Moore (B)	Masiero (C)[a]	Current study
Detection rate[b]	53.3% Noise	59.4%/59.4%/48% Noise/pulses/tones	87.5%/71.4%/73.7% Noise/speech/music	See Table II
Critical det. rate[c]	52.33% (1800 trials, 1 test)	59.38% (192 trials, 6 tests)	55.4%/55.19%/55.65% (208–241 trials, 3 tests)	Assessed on participant basis
Audio content	Noise, 0.5–16 kHz, varying spectral shape	Noise, 0.12–15 kHz Pulse trains, 0.1–15 kHz Complex tone, 0.1–4.6 kHz	Noise, 0.2–20 kHz Speech, 0.2–8 kHz Music, 0.2–10 kHz	Noise, 0.1–16.4 kHz, Speech, 0.1–16.4 kHz
Test Environment	Static synthesis Extra-aural headphones Anechoic 6 sources, (Around listener)	Static synthesis CTC loudspeakers Anechoic 1 source, (Frontal)	Static synthesis circumaural headphones Anechoic 24 sources, (Around listener)	Dynamic synthesis Extra-aural headphones Anechoic and reverberant 2 sources, (Frontal and lateral)
Test method	4I/2AFC Listening once With training With feedback 6 participants, (Experienced)	4I/2AFC Listening once With training With feedback 8 participants, (Mostly experienced)	3I/3AFC Listening three times Without training Without feedback 40 participants, (Unexperienced)	3I/2AFC Unlimited listening With training Without feedback 9 participants, (Experienced)

[a] Detection rates from *D* (experiment with blocked ear canal measurements): 79.3% (noise), 66.8% (speech), 69.8% (music). Critical detection rate: 52.75% (800 trials, 3 tests).

[b] Averaged across participants and sources; detection rates from *C*, and *D* were transformed to 2AFC detection rates.

[c] Dunn-Sidák correction for multiple testing was applied to the initial type 1 error level of 5%.

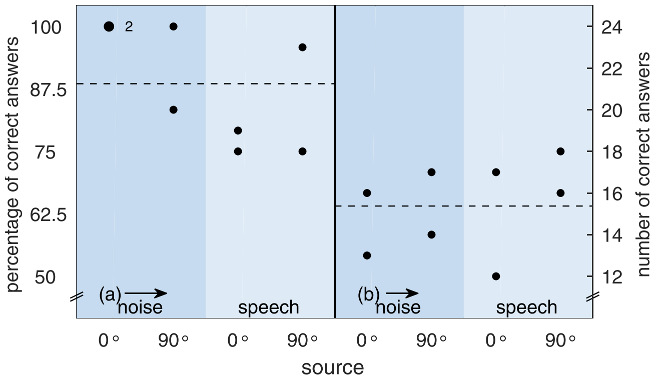

FIG. 6. (Color online) The 2AFC detection rates for two participants and the anechoic environment. (a) Results for dynamic binaural synthesis and listening at will; (b) results for static synthesis and listening only once. The size of the dots and the numbers next to them indicates how many participants scored identical results. Dashed lines show averaged detection rates.

in *A* and *B*, compared to listening up to three times in *C* and *D*, or unlimited listening as allowed in our study, where the subjects could listen as often as they wanted, in any order, and switch between stimuli at any time. To investigate the extent to which this difference increased the sensitivity of the test, we conducted a 3I/2AFC test in the anechoic environment and sequentially presented the stimuli only once. Two participants, who also took part in the previous test a couple of days earlier, were selected to allow for a direct comparison of their detection rates under both conditions. As shown in Fig. 6, the changed mode of presentation caused a considerable and statistically significant decrease in detection rates from 89% for unrestricted listening to 64% for restricted listening (single sided Wilcoxon signed rank test for dependent samples, paired across audio contents and source positions, $p < 0.01$). The detection rates were now similar for noise (62.5%) and speech (65.6%), and they are well comparable to the detection rates of earlier studies. These results emphasize the impact of the test design and point out that a direct comparison of detection rates across studies has to be carried out with caution.

Interestingly, neither *A* nor *C* and *D* reported notable effects of the source position. In the current study, listeners could change their head orientation and were thus not restricted to a fixed relative source position anyway. Also in this case, the two source positions provided (frontal/lateral) did not entail significantly different detection rates.

V. CONCLUSION

In the present study, we assessed whether the binaural re-synthesis of electro acoustic sources can be discriminated from the corresponding real sound field under optimal test conditions, such as individual BRIRs and no headphone repositioning, as well as state-of-the-art measurement, post-processing, and rendering. For the first time, perceptual "authenticity" was tested for spatial environments with different degrees of reverberation, and dynamic binaural simulations accounting for azimuthal head-movements of the listeners. Remaining differences were evaluated with respect to the acoustical signal as well as with a finely differentiated inventory of perceptual attributes.

For testing the perceptual authenticity at the level of individual participants, we conducted a minimum effect size N-AFC listening test with balanced type one and type two errors. In order to maximize the sensitivity of the test, we allowed for repeated listening and switching between the stimuli, and provided audio content suitable to uncover different potential flaws of the simulation. The influence of the acoustical environment (anechoic/dry/wet) and the source position (frontal/lateral) were analyzed as independent variables.

In agreement with earlier studies, we found that—even with these prerequisites—for a pulsed pink noise sample all participants could reliably detect differences between reality and simulation. For the speech sample, however, the detection rates of individual participants ranged from 62% to 91% with a mean of 76%. Hence, for almost half of the trained expert listeners, who could immediately and repeatedly compare the two stimuli under optimal listening conditions, the simulation can be considered as perceptually authentic.

An interaction between audio content and room was observed, with detection rates being highest for the anechoic environment and lowest for the reverberation chamber in case of the speech content. The remaining differences between simulation and reality, that stem from measurement uncertainties, manifest themselves mainly in a degradation of tone color related qualities rather than in localization or spatial impression. This in turn explains why pink noise with its broadband spectral content provides the strongest cues to identify these differences.

In both analyses (technical and the perceptual), the anechoic condition proved to be the worst case for binaural re-synthesis. With increasing reverberation time and constant source-receiver distances, i.e., with decreasing direct-to-reverberant energy ratios (DRRs), the spectral differences between reality and simulation are partially smoothed out, corresponding to lower detection rates for the reverberant environments compared to the anechoic situation. The difference between the frontal and lateral source position, on the other hand, had no significant influence for any of the three spatial environments.

The results suggest that for "everyday audio content" with limited spectral bandwidth such as speech and music, an authentic virtual representation of acoustic environments can be achieved by using individual dynamic binaural synthesis, if sufficient care is taken for the acquisition, post-processing and rendering of the corresponding binaural impulse response datasets. While natural acoustic sources with their time-variant behavior present particular challenges, this should always be possible for loudspeakers and electro-acoustic reproduction systems and enable their perceptual evaluation by binaural re-synthesis without a relevant loss in quality.

ACKNOWLEDGMENTS

This work was funded by the German Research Foundation (DFG WE 4057/3-1). We would like to thank the participants from the listening test for their patience and commitment.

[1] J. Blauert, *Spatial Hearing. The Psychophysics of Human Sound Localization*, revised edition (MIT Press, Camebridge, MA,1997), 495 pp.

[2] H. Møller, "Fundamentals of binaural technology," Appl. Acoust. **36**, 171–218 (1992).

[3] F. L. Wightman and D. J. Kistler, "Headphone simulation of free field listening. I: Stimulus synthesis," J. Acoust. Soc. Am. **85**, 858–867 (1989).

[4] W. R. Thurlow, J. W. Mangels, and P. S. Runge, "Head movements during sound localization," J. Acoust. Soc. Am. **42**, 489–493 (1967).

[5] K. I. McAnally and R. L. Martin, "Sound localization with head movement: Implications for 3-d audio displays," Front. Neurosci. **8**, 1–6 (2014).

[6] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. Katz, and C. de Boishéraud, "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," J. Acoust. Soc. Am. **141**, 2011–2023 (2017).

[7] C. Kim, R. Mason, and T. Brookes, "Head movements made by listeners in experimental and real-life listening activities," J. Audio Eng. Soc. **61**, 425–438 (2013).

[8] D. S. Brungart, B. D. Simpson, and A. J. Kordik, "The detectability of headtracker latency in virtual audio displays," in *Eleventh Meeting of the International Conference on Auditory Display (ICAD)* (Limmerick, Ireland, 2005), pp. 37–42.

[9] A. Lindau and S. Weinzierl, "On the spatial resolution of virtual acoustic environments for head movements on horizontal, vertical and lateral direction," in *EAA Symposium on Auralization* (Espoo, Finland, 2009), 6 pp.

[10] E. H. A. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," J. Acoust. Soc. Am. **107**, 528–537 (2000).

[11] A. Lindau, J. Estrella, and S. Weinzierl, "Individualization of dynamic binaural synthesis by real time manipulation of the ITD," in *128th AES Convention, Convention Paper* (London, UK, 2010), 10 pp.

[12] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments," Acta Acust. Acust. **98**, 804–810 (2012).

[13] C. Pike, F. Melchior, and T. Tew, "Assessing the plausibility of non-individualised dynamic binaural synthesis in a small room," in *AES 55th International Conference* (Helsinki, Finland, 2014), 8 pp.

[14] A. H. Moore, A. I. Tew, and R. Nicol, "An initial validation of individualised crosstalk cancellation filters for binaural perceptual experiments," J. Audio Eng. Soc. **58**, 36–45 (2010).

[15] B. Masiero, "Individualized binaural technology. measurement, equalization and perceptual evaluation," Doctoral thesis, RWTH Aachen, Aachen, Germany (2012), pp. 99–115.

[16] J. Oberem, B. Masiero, and J. Fels, "Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods," Appl. Acoust. **114**, 71–78 (2016).

[17] F. Wightman and D. Kistler, "Headphone simulation of free field listening. II: Psychological validation," J. Acoust. Soc. Am. **85**, 868–878 (1989).

[18] T. Hiekkanen, A. Mäkivirta, and M. Karjalainen, "Virtualized listening tests for loudspeakers," J. Audio Eng. Soc. **57**, 237–251 (2009).

[19] H. Wierstorf, "Perceptual assessment of sound field synthesis," Doctoral thesis, Technische Universität Berlin, Berlin, Germany (2014), Chap. 4.

[20] M. Frank, "Phantom sources using multiple loudspeakers in the horizontal plane," Doctoral thesis, University of Music and Performing Arts, Graz, Austria (2013), Chap. 2.

[21] A. Lindau and F. Brinkmann, "Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings," J. Audio Eng. Soc. **60**, 54–62 (2012).

[22] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer characteristics of headphones measured on human ears," J. Audio Eng. Soc. **43**, 203–217 (1995).

[23] M. Paquier and V. Koehl, "Discriminability of the placement of supra-aural and circumaural headphones," Appl. Acoust. **93**, 130–139 (2015).

[24] A. H. Moore, A. I. Tew, and R. Nicol, "Headphone transparification: A novel method for investigating the externalisation of binaural sounds," in *123rd AES Convention* (New York, 2007), Convention Paper 7166, 9 pp.

[25] F. Brinkmann, A. Lindau, M. Vrhovnik, and S. Weinzierl, "Assessing the authenticity of individual dynamic binaural synthesis," in *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics* (Berlin, Germany, 2014), pp. 62–68.

[26] S. G. Norcross, M. Bouchard, and G. A. Soulodre, "Inverse filtering design using a minimal phase target function from regularization," in *121th AES Convention* (San Francisco, CA, 2006), Convention Paper 6929, 8 pp.

[27] A. Andreopoulou, D. R. Begault, and B. F. G. Katz, "Inter-laboratory round robin HRTF measurement comparison," IEEE J. Sel. Topics Signal Process. **9**, 895–906 (2015).

[28] J. G. Tylka, R. Sridhar, and E. Y. Choueiri, "A database of loudspeaker polar radiation measurements," in *139th AES Convention* (New York, 2015), e-Brief 230, 4 pp.

[29] V. Erbes, F. Schultz, A. Lindau, and S. Weinzierl, "An extraaural headphone system for optimized binaural reproduction," in *Fortschritte der Akustik – DAGA 2012* (Darmstadt, Germany, 2012), pp. 313–314.

[30] A. W. Mills, "On the minimum audible angle," J. Acoust. Soc. Am. **30**, 237–246 (1958).

[31] F. Brinkmann and S. Weinzierl, "AKtools–an open toolbox for acoustic signal acquisition, processing, and inspection," www.ak.tu-berlin.de/AKtools (Last viewed June 2016).

[32] S. Müller and P. Massarani, "Transfer function measurement with sweeps. directors cut including previously unreleased material and some corrections," J. Audio Eng. Soc. **49**, 443–471 (2001).

[33] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida, "Mechanism for generating peaks and notches of head-related transfer functions in the median plane," J. Acoust. Soc. Am. **132**, 3832–3841 (2012).

[34] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkmann, and S. Weinzierl, "A Spatial Audio Quality Inventory (SAQI)," Acta Acust. Acust. **100**, 984–994 (2014).

[35] S. Ciba, F. Brinkmann, and A. Lindau, "WhisPER. A MATLAB toolbox for performing quantitative and qualitative listening tests [computer software] (version 1.9.0)," http://dx.doi.org/10.14279/depositonce-31.2 (2014).

[36] A. Lindau, T. Hohn, and S. Weinzierl, "Binaural resynthesis for comparative studies of acoustical environments," in *122th AES Convention* (Vienna, Austria, 2007), Convention Paper 7032, 10 pp.

[37] L. Leventhal, "Type 1 and type 2 errors in the statistical analysis of listening tests," J. Audio Eng. Soc. **34**, 437–453 (1986).

[38] K. R. Murphy and B. Myors, "Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model," J. Appl. Psychol. **84**, 234–248 (1999).

[39] D. Pralong and S. Carlile, "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space," J. Acoust. Soc. Am. **100**, 3785–3793 (1996).

[40] B. C. J. Moore, S. R. Oldfield, and G. J. Dooley, "Detection and discrimination of spectral peaks and notches at 1 and 8 kHz," J. Acoust. Soc. Am. **85**, 820–836 (1989).

[41] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," J. Acoust. Soc. Am. **109**, 1110–1122 (2001).

[42] F. Brinkmann, R. Roden, A. Lindau, and S. Weinzierl, "Audibility and interpolation of head-above-torso orientation in binaural technology," IEEE J. Sel. Topics Signal Process. **9**, 931–942 (2015).

[43] C. Ryan and D. Furlong, "Effects of headphone placement on headphone equalisation for binaural reproduction," in *98th AES Convention* (Paris, France, 1995).

[44] K. A. J. Riederer, "Part Va: Effect of head movements on measured head-related transfer functions," in *18th International Congress on Acoustics* (Kyoto, Japan, 2004), pp. 795–798.