Acta Acustica 2021, **5**, 8

https://doi.org/10.1051/aacus/2020034

Available online at:
https://acta-acustica.edpsciences.org

SCIENTIFIC ARTICLE OPEN ACCESS

Toward realistic binaural auralizations – perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario

Matthias Blau[1,2,*], Armin Budnik[1], Mina Fallahi[1], Henning Steffens[2,3], Stephan D. Ewert[2,3], and Steven van de Par[2,4]

[1] Institut für Hörtechnik und Audiologie, Jade Hochschule, Ofener Str. 16, 26121 Oldenburg, Germany
[2] Cluster of Excellence "Hearing4All"
[3] Medizinische Physik, Carl von Ossietzky Universität Oldenburg, Carl-von-Ossietzky-Str. 9-11, 26129 Oldenburg, Germany
[4] Acoustics Group, Carl von Ossietzky Universität Oldenburg, Carl-von-Ossietzky-Str. 9-11, 26129 Oldenburg, Germany

Received 28 June 2020, Accepted 21 December 2020

Abstract – In order to make full use of their potential to replace experiments in real rooms, auralizations must be as realistic as possible. Recently, it has been shown that for speech, head-tracked binaural auralizations based on *measured* binaural room impulse responses (BRIRs) can be so realistic, that they become indistinguishable (or nearly so) from the real room [1, 2]. In the present contribution, perceptual comparisons between the auralized and the real room are reported for auralizations based both on *measured* and *simulated* BRIRs. In the experiment, subjects sitting in the real room rated the agreement between the real and the auralized room with respect to a number of attributes. The results indicate that for most attributes, the agreement between the auralized and the real room can be very convincing (better than 7.5 on a nine-point scale). This was not only observed for auralizations based on measured BRIRs, but also for those based on simulated BRIRs. In the scenario considered here, the use of individual head-related impulse responses (HRIRs) does not seem to offer any benefit over using HRIRs from a head-and-torso-simulator.

Keywords: Binaural auralization, Classroom acoustics, Computer simulation of acoustics in enclosurs, Perceptual evaluation, Head-related impulse responses

1 Introduction

Auralization, i.e. making sound fields audible by rendering them via headphones or loudspeakers, has been used for many years, see [3] for an early overview. Despite this long history, the use of auralization as a replacement for listening in a real room has not become widespread. A majority of acoustical consultants who participated in a recent survey indicated to use it mainly for marketing and communication purposes – more than 50%, compared to 34% using it to "test ideas" or as a "verification tool" [4]. Yet, auralization has much more to offer: It has the potential to replace experiments in real rooms, with applications in architectural acoustics, psychology, therapeutic training, etc. [5]. This requires, however, the auralization to be as realistic as possible.

Previous studies regarding the realism of binaural (i.e., rendered via headphones) auralizations have mainly focused on comparing auralizations based on simulated sound fields to those based on measured sound fields, see e.g. [6–10]. The latent (and sometimes explicitly formulated) assumption underlying the cited studies is that measurements represent the "reality" and that simulations could also represent the reality if they were indistinguishable from measurements in the direct comparison.

This assumption implicitly assumes that listening to auralizations based on measured (and, if there were successful simulations, on simulated) binaural room impulse responses (BRIRs) would create the same auditory perception as listening to the real source(s) in the real room. Even if difficulties in exactly reproducing natural sources are put aside by using loudspeakers as sources, one must acknowledge that this remains an uncertain assumption.

To begin with, the measurement of individual BRIRs (which must be performed for all conceivable head-above-torso orientations in dynamic scenarios, see e.g. [1]) is challenging and prone to inaccuracies, e.g. due to lacking control of head position/orientation, let alone that such measurements are extremely exhausting for the subjects participating. But even if a rendering was perfectly indistinguishable from the original presentation in the real room, it might create severe artifacts such as lacking externalization when presented in a different room, because the room that is rendered may not correspond to the room in which the

*Corresponding author: matthias.blau@jade-hs.de

rendering is reproduced. The perception in the reproduction room will be influenced by other modalities such as visual information, which may provide conflicting expectations with respect to the real room. This has recently become known as the "room divergence effect" [11]. Hence, in order to establish realistic auralizations, it is not sufficient to rely on perceptual comparisons performed in either the real or the reproduction room *alone*, although such investigations are valuable first steps. An important consequence is that the frequently used ABX tests, while being very accurate and reliable, may not be valid to characterize the perception transfer to the reproduction room.

A few investigations were concerned with the *direct comparison* between a presentation of real sources (loudspeakers) and binaural auralizations *in the real room*. In [12], static binaural auralizations were compared to loudspeaker presentations in an anechoic environment. The binaural auralizations were found to be plausible, i.e. subjects could not reliably detect whether a given presentation originated from the loudspeaker or from the auralization presented via headphones. However, in direct comparisons differences between loudspeaker presentation and auralization could be detected for noise pulses and, to a lesser degree, for music and speech.

In [1], a blind comparison between head-tracked auralizations based on individually measured BRIRs (and rendered on a custom-made acoustically open headphone resting on the subject's head throughout the experiment) and the original loudspeaker presentation in the room in question was performed for three different rooms. It was shown that the subjects could reliably detect differences between the auralization and the real room for noise stimuli, but not for speech stimuli. Hence, a realistic auralization of speech, based on measured BRIRs, is possible.

This was confirmed in our previous study [2], in which head-tracked binaural auralizations of speech stimuli, presented by loudspeakers in a typical lecture room, were compared to the same stimuli reproduced by the loudspeakers in the real room, by repeatedly switching between headphone and loudspeaker presentation. More specifically, for auralizations based on *measured* head-and-torso-simulator (HATS) BRIRs, the agreement to the real room was rated with a median of close to "excellent" (better than eight on a nine-point scale) for most attributes considered in that study.

The next step would then be to compare auralizations based on *simulated* BRIRs to the original presentation in the real room. This was also part of the investigations in [2], in which BRIRs were simulated using the RAZR room simulation package [13] with different sets of head-related impulse responses (HRIRs) to represent listener directivities. These HRIR sets included measured head-and-torso-simulator (HATS) or individual HRIRs, both of which where in addition rendered via the virtual artificial head (VAH) technology [14, 15]. It was found that ratings regarding the agreement with the real room were significantly lower for auralizations based on the simulated compared to auralizations based on measured BRIRs, with medians between "fair" and "good" (five and seven on the nine-point scale).

A careful re-analysis of the renderings and the results of [2] led us to believe that the most salient weakness of the room simulation package that was used in these experiments was its limitation to simulate omnidirectional sources only. Since the directionality of loudspeakers and humans tends to be stronger at high frequencies, the usage of omnidirectional sources in the simulation leads e.g. to more reverberant energy at high frequencies as compared to what would occur with sources with strong directionality at high frequencies. This is in line with findigs from e.g. [16, 17] showing that the inclusion of source directivity improves the perceived realism of auralizations.

In an extended version of the simulation package [18], the limitation to omnidirectional sources was removed and the question arose again to which extent various BRIR simulation methods are able to provide a realistic auralization of a real room, specifically in comparison to auralizations based on measured BRIRs. The following specific questions are addressed: (1) How close can the auralization (be it based on measured or simulated BRIRs) come to the real room, (2) What is the effect of using simulated instead of measured BRIRs in the simulations, (3) What is the effect of using HATS versus individual HRIRs in the simulations, (4) What is the effect of ignoring the source directivity in the simulations, and (5) What is the effect of using HRIRs synthesized with the VAH technology instead of the original ones.

2 Methods

2.1 Room configuration

The room under investigation was a small lecture room at Jade Hochschule (dimensions 7.12 × 11.94 × 2.98 m^3). The ceiling was completely fitted with broad-band absorbing perforated plaster board, backed by a 5 cm layer of mineral wool and a rearward air space of about 30 cm. The floor was linoleum covered, window curtains were open, and typical equipment such as wooden chairs and desks were in the room, see also Figure 1.

Subjects evaluating the auralizations sat in the third row, approximately 0.65 m to the right of the room center axis. The position of the chair on which the subjects sat was controlled, no other measures to control position were taken. Subjects could freely rotate their heads but were not explicitly encouraged to do so. The height of the ear axis above the floor was assumed to be 1.3 m. This was also used for the measurement respectively the simulation of BRIRs.

Three two-way active loudspeakers (two Genelec model 8030b, one Genelec model 8030c, Genelec Oy, Iisalmi, Finland) were used to represent three different source positions. The frontal loudspeaker (marked as "S1" in Fig. 1) was used to represent a lecturer at 4.8 m in front of the subject. Since this speaker represented a standing person, its height (defined by the crossing point of the two diagonals of the loudspeaker box front surface) was a little higher than that of the other two loudspeakers deemed to represent fellow students (1.6 m vs. 1.3 m). The left loudspeaker (marked as "S2" in Fig. 1) was located in the same

Figure 1. Room configuration: lecture room at Jade Hochschule, with three loudspeakers (labeled S1, S2, and S3), and one subject sitting in the third row. The screen displaying the user interface for the perceptual evaluation was placed directly onto the blackboard in front of the subject.

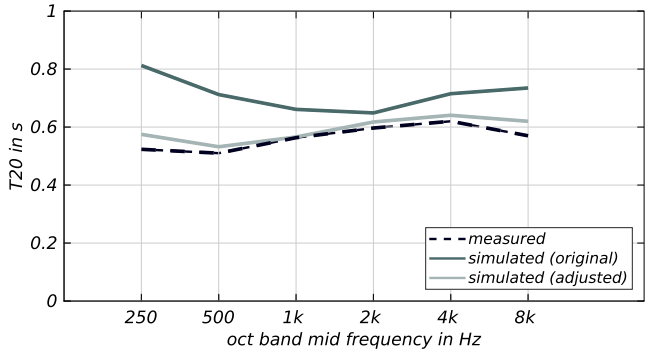

Figure 2. Measured and simulated reverberation times at the subject's position, averaged over the three loudspeaker positions. The dashed line represents measured values (using a typical measurement microphone). The solid lines represent simulated values using the RAZR simulation package, with absorption coefficients originally estimated from inspection of the walls (darker line) and adjusted absorption coefficients (lighter line).

row as the subject, at a distance of 3.9 m to the left. The rear loudspeaker (marked as "S3" in Fig. 1) was at the right side behind the subject at a distance of 1.8 m. The corresponding azimuth and elevation angles were (0°, 4°) for S1, (90°, 0°) for S2 and (245°, 0°) for S3. All loudspeakers had their center axis azimuths directed towards the subject with 0° elevation.

In order to characterize the room at the subject's position acoustically, room impulse respones with respect to the three loudspeakers were measured using a measurement microphone (type 40AF, GRAS Sound & Vibration A/S, Holte, Denmark). Resulting values for T_{20}, averaged over the three loudspeaker positions, are shown in Figure 2 (dashed line). Octave band values of T_{20} are on average slightly lower than 0.6 s, with a broad minimum around 500 Hz, caused by the acoustically treated ceiling.

2.2 BRIR sets used in the auralizations

Six different BRIR sets were considered in the present study.

The first BRIR set (labeled "meas: HATS" in the following) consisted of BRIRs measured in the room under question with a commercial HATS (KEMAR type 45BB, GRAS Sound & Vibration A/S, Holte, Denmark) for 37 azimuthal head-above-torso orientations (−90° to +90° in 5° steps), as described in [2]. The HATS was fitted with foam earplugs into which a 3D-printed support carrying a MEMS microphone (Knowles type SPV0840LR5H, Knowles Electronics, LLC, Itasca, IL, USA) was assembled such that a blocked ear condition was obtained, see [19] for details. A multiple exponential sweep stimulus [20], with modifications proposed in [21], from 20 Hz to the Nyquist frequency in about 20 s, shifted by 4 s for subsequent loudspeakers, was used. Stimulus generation, sound playback/recording and impulse response computation were realized on a laptop computer equipped with an external audio interface (RME Fireface UC, Audio AG Haimhausen, Germany) operated at $F_s = 44.1$ kHz, using custom Matlab (The MathWorks, Inc., Natick, MA, USA) and Pure Data [22] scripts. Impulse responses were cut to a length of about 410 ms, corresponding to the length of the simulated BRIRs. In all cutting operations in the present work, a 50 point half Hann window was used to fade out the respective impulse responses. The chosen length of the impulse responses enabled a dynamic range of more than 40 dB, which matched or exceeded the usable dynamic range in the listening test, see Section 2.3.

The remaining five BRIR sets were obtained from room simulations using a preliminary, improved version of the RAZR package [13, 18], in which the source directivity could be taken into account. The source directivity of the loudspeakers was first measured in an anechoic room using a free-field-equalized microphone (type 40AF, GRAS Sound & Vibration A/S, Holte, Denmark) at 1 m distance, with a spatial resolution of 5° in both azimuth and elevation. The measured directivity data were compiled as set of impulse reponses truncated at about 227 ms (10 000 points at 44.1 kHz sampling rate). The measured loudspeaker directivity data was used both in the simulation of early reflections (by the image source method, ISM) and implicitly also in that of the late reverberation in RAZR. Within the ISM, the orientations of the direct sound and the image sources relative to the receiver were used to select the corresponding measured loudspeaker impulse response using the nearest neighbor method. Here, these calculations were performed up to the third order of image sources, which was considered sufficient based on informal listening tests with two subjects. For the late reverberation a feedback delay network (FDN) is used in RAZR. The input to the FDN are the last order reflections (third order in this case) calculated in the ISM. Effects of directivity-dependent spectral filtering as present in these reflections are thus contained in the FDN input and affect the average spectral coloration of the FDN output, see [18] for more details. It is assumed that this average spectral representation of the effect of source directivity in the late reverberant tail

is sufficient to obtain a perceptually accurate auralization in combination with accurate source directivity modelling in the direct sound and the early reflections.

In the room simulation, RAZR uses a strongly simplified room model: the room geometry is approximated by a shoebox shape and uniform absorption coefficients (in octave bands) per room wall are used. In order to achieve realistically sounding simulations, the absorption coefficients of the six room walls were manually adjusted such that the simulated reverberation time was within 5% of the measured (using a typical measurement microphone, see Sect. 2.1 above) octave-band T_{20}, averaged over the three loudspeaker positions (see Fig. 2), by preserving the general frequency-dependent character of absorption of each wall (e.g. broad-band absorption centered around 500 Hz for the ceiling).

All simulated BRIR sets were computed for 333 head orientations (the 37 azimuthal orientations that have already been used with the HATS, times nine elevations from −30° to +30° in 7.5° steps). They differed in the way the room simulations were converted to binaural impulse reponses by applying different sets of head-related impulse responses (HRIRs). In any case, HRIR sets had an azimuthal resolution of 5° with twelve elevations (−30° to +30° in 7.5° steps plus +45°, +60° and +75°), i.e., a total of 864 directions of incidence were used in all HRIR sets. The binaural impulse reponses were obtained by convolving the directional impulse reponses obtained from the room simulation with the HRIRs matching the respective direction of incidence, again using the nearest neighbor method.

For the second BRIR set (labeled "sim: HATS"), measured HRIRs of the HATS were used, whereas for the third one (labeled "sim: indivHRIRs"), measured HRIRs of the individual subjects were used. Both individual and HATS HRIR sets were measured in a custom-made rig, as described in [2]. The rig consisted of 24 small active loudspeakers (Speedlink type SL-8902-GY Xilu, Jöllenbeck GmbH, Weertzen, Germany) mounted on an aluminum arc (radius 1.25 m) that rotated around the subject in 5° steps. Loudspeakers were individually equalized in magnitude and phase. The rig was located in another lecture room of comparable size and acoustics. Floor and ceiling were equipped with absorbent foam wedges of 68 cm length and room reflections were removed by truncating the impulse responses to a length of about 6 ms (256 points at 44.1 kHz). Measurement procedures were the same as the ones described above for the BRIR measurements, with the exception that now 24 loudspeakers were used and the stimulus length was slightly modified (sweep from 100 Hz to Nyquist in 10 s, shifted by 0.3 s for subsequent loudspeakers). The low frequency limit of the HRTFs was approximately 300 Hz, mainly due to the characteristics of the miniature loudspeakers. Subjects were seated in the center of the arc on a chair with small back and head rests. Their exact position was adjusted vertically and horizontally such that the ear axis passed through the center of the arc and was perpendicular to the direction of frontal incidence, and the intra-aural center coincided with the center of the arc.

For the fourth BRIR set (labeled "sim: indivHRIRs*"), the individually measured HRIRs were recreated using the VAH technology: More specifically, a 32-channel three-dimensional microphone array (array extension $11 \times 11 \times 7$ cm^3), for which the steering vectors were simulated, was used together with filter coefficients for each of the microphones to obtain synthesized HRIRs at the desired directions. The filter coefficients in turn were obtained by minimizing, for left and right ear separately, in each frequency band, the squared absolute differences between the original directivity pattern (individually measured head related transfer function in that frequency band at 72 equidistant horizontal source positions, pre-processed according to [23]) and that of the filter-and-sum beamformer, summed over the 72 positions, subject to constraints on the mean white noise gain (mWNG) [24] and on the spectral distortion at the 72 equidistant horizontal source positions [15]. As a result of this procedure, horizontal HRTFs could be synthesized with a monaural spectral deviation between −1.5 dB and +0.5 dB and a mWNG of at least 0 dB up to 8 kHz. For non-horizontal directions the spectral distortion was much higher since these directions were not included in the calculation of the filter coefficients. The re-synthesized HRTFs (for the same directions as all other HRIR sets) were then converted to corresponding impulse responses and processed in the same manner as the other HRIR sets.

For the fifth BRIR set (labeled "sim: HATS omni"), measured HATS HRIRs were again used, but this time, the source directivities were ignored in the simulation.

Finally, analytical rigid sphere HRIRs [25] were used as anchor (sixth BRIR set, labeled "sim: sphere"), assuming "ear" positions at ±100° on the equator of a sphere of $d = 16$ cm diameter. This anchor was choosen because it provides a simple and reproducible approximation of the real spatial and spectral cues perceived by human listeners except for substantially differing in terms of a full lack of torso- and pinna-related cues. As with the "sim: HATS omni" condition the "sim: sphere" BRIR sets were computed for omnidirectional sources.

For all except the "sim: sphere" BRIR set, headphone equalization (in amplitude and phase) was used. The equalization filters were based on individually measured headphone transfer functions (HPTFs). The HPTF measurements were performed immediately after the HRIR measurement, with the ear microphones still in place. The subjects took off and on the headphones up to nine times. Out of these measurements, the one with the shallowest dips in the 8…12 kHz range was used for inversion [14]. The corresponding impulse responses were truncated to about 93 ms (4096 samples at 44.1 kHz) and inverted using the method from [26] with a regularization parameter of $\beta = 10$ times the average mean square value of both headphone impulse responses.

The loudspeaker frequency response is implicitly contained in the first four BRIR sets which rely on measured loudspeaker directivities. The "sim: HATS omni" and "sim:sphere" BRIR sets, on the other hand, had to be equalized to account for the loudspeaker frequency

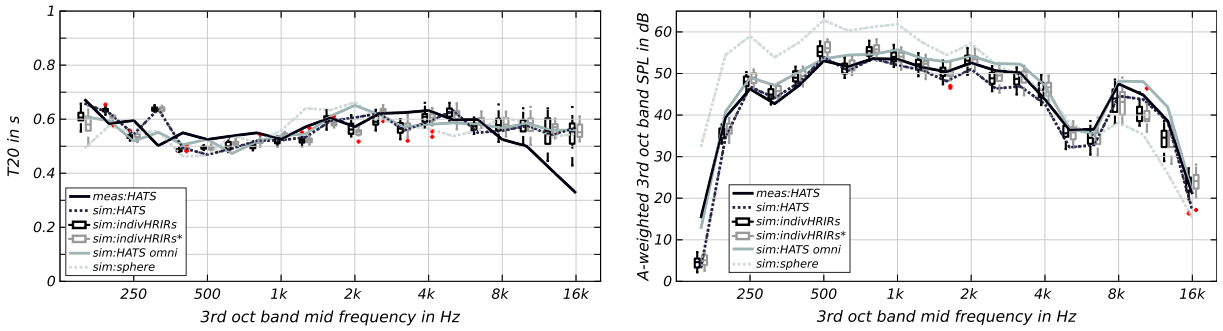

Figure 3. Objective measures, derived from BRIRs for frontal head orientation (averaged over all three loudspeaker positions and left and right ear), for all BRIR sets considered in this study. Box plots refer to BRIR sets of individual subjects, lines to non-individual BRIR sets. **Left:** resulting reverberation times (T20), **right:** resulting A-weighted third-octave band sound pressure levels, obtained after convolution of the stimulus with the respective BRIRs.

response. To this end, equalization filters were designed (one for each loudspeaker) by matching the "sim: HATS omni" simulation to the "meas: HATS" BRIR set, for the initial head orientation (0° azimuth and 0° elevation), and after having spectrally smoothed them into octave bands using the method proposed in [27]. The same equalization filters were then also applied to the "sim: sphere" BRIR set.

In Figure 3, averaged (over loudspeakers and ears) third-octave band spectra of T_{20} and the A-weighted sound pressure level for all BRIR sets used in this study are shown. The latter were obtained by convolving the stimulus used in the listening test (female speech, see Sect. 2.3 below) with the BRIR sets for frontal head orientation, averaging over speakers and left and right ears and scaling the absolute spectrum such that the A-weighted total level produced by speaker 1 with the "meas: HATS" BRIR set matched the one measured at the listener position using a commercial sound level meter (type 2260, Brüel & Kjær A/S, Nærum, Denmark).

Regarding reverberation times, it can be seen that T_{20} is very similar for all BRIR sets, although in contrast to Figure 2, they are now based on BRIRs (for frontal head orientation, average of left and right ear) instead on monaural impulse responses and are shown in third-octave instead of octave bands. Only at frequencies above 8 kHz a noteworthy discrepancy is observed: The values based on measured HATS BRIRs decrease towards higher frequencies whereas the values based on simulated BRIRs do not. This can be explained by the fact that the absorption filters in RAZR have only been optimized in the default 250 Hz to 8 kHz octave bands (see Fig. 2). In the interest of computational efficiency, the filters remain constant outside that frequency range and accordingly the simulated reverberation times remain approximately constant at frequencies above 8 kHz. The simulations can be even further improved by additionally fitting the 16 kHz octave band at the expense of slightly less efficient filters.

Regarding the resulting spectra, one can observe that the anchor ("sim: sphere" BRIR set) is markedly different from all other BRIR sets, by emphasizing low and mid frequencies up to 2 kHz and attenuating high frequencies above 8 kHz in comparison. This resulted in a strong coloration of the anchor condition. All other BRIR sets were similar in that the spectral differences between them were comparable to the interindividual differences in the BRIR sets for individual subjects, with only a few exceptions: First, at 125 Hz, the "meas: HATS" and "sim: HATS omni" BRIR sets were 10 dB above the "sim: HATS", "sim: indivHRIRs" and "sim: indivHRIRs*" BRIR sets. Since the absolute (A-weighted) level in this frequency band was about 40 dB below that in the mid-frequency bands, this can probably be ignored. Second, the non-individual BRIR sets, in particular "sim: HATS omni", exhibited somewhat higher levels than the individual BRIR sets in the 10 and 12.5 kHz bands.

2.3 Listening test

The listening test is based on [2]. It was designed as comparison between different auralized (head-tracked) binaural reproductions and the reference reproduction via loudspeakers of a speech signal, while the subject was sitting in the real room, see Figure 1. A dry recording of "Nordwind und Sonne" (text version from the IPA Handbook 1999 [28], first sentence), spoken by a female speaker in an acoustically treated lecture room, was used as stimulus. As in [2], the audio signal was high-pass filtered using a digital 6th order Butterworth filter with 200 Hz cutoff frequency. The A-weighted sound pressure level at the listener position was 60 dB when speaker 1 was used as source. The A-weighted background noise level in the room was between 20 and 25 dB, depending on outdoor conditions.

For the comparison, the subject was taking the headphones (model HD 800, Sennheiser electronic GmbH & Co. KG, Wedemark, Germany) off and on and simultaneously switched between headphone and loudspeaker reproduction. In order to switch from headphone to loudspeaker reproduction when taking off the headphones, the subject pushed and held down a button attached to the headphones, resulting in fading out the headphone reproduction (fade-out time 50 ms) and fading in the loudspeaker reproduction (fade-in time 400 ms). Conversely, when the subject put the headphones on and released the button,

the loudspeaker reproduction was faded out and the headphone reproduction faded in.

As a consequence of this procedure, the reference (loudspeaker reproduction) was not hidden. This is different from [12] and [1] where subjects wore headphones which were claimed to be acoustically open throughout the experiment. The advantage of headphones being worn throughout the experiment and the HRTF measurement is that one can hide the reference condition and that artifacts due to the repositioning of the headphones can be avoided. On the other hand, listening to an acoustical scene created by external sources in a room is different from listening to the same scene with the ears being occluded by devices, even if the latter are supposed to be acoustically open: The (direction-dependent) sound scattering caused by the device can be thought of as acting like modified pinnae – the effect can certainly be captured if HRTFs are measured with the device on and one can perceptually adapt to them, but it is not the original situation that one wants to reproduce.

The head-tracked binaural renderings were computed in real-time, using a custom software based on the time-variant partitioned convolution algorithm [29], which is publicly available at https://github.com/TGM-Oldenburg/TVOLAP.

Twenty subjects (11 male, 9 female, average age 29.1 years, 125 Hz to 8 kHz pure-tone thresholds better than 15 dB HL) participated in the study. Ten of them had substantial experience in listening tests, four were naive. As in [2], their task was to rate how much the headphone reproduction agreed to the loudspeaker reproduction with respect to selected attributes.

Based on informal listening tests with four subjects, five attributes were identified to be critical with respect to the comparison between real source and auralization in the given setup. These attributes were "Halligkeit" (reverberance), "Breite der Quelle" (source width), "Entfernung der Quelle" (source distance), "Richtung der Quelle" (source direction), and "Klangverfärbungen" (coloration). In order to cover aspects that were not explicitly contained in any of these attributes, the attribute "Gesamtqualtität" (overall quality) was also considered. Since we felt that the original selection of attributes already covered the majority of all possible perceptive differences between loudspeaker presentation and auralizations, and at the same time to limit the number of ratings to be given by the subjects, we decided to omit a solicitation of ratings for "Klangverfärbungen" (coloration) and instead have this attribute included in "Gesamtqualtität" (overall quality). In order to ensure that "Gesamtqualtität" (overall quality) indeed referred to everything not yet covered by the other attributes, it was rated as last attribute by all subjects and subjects were explicitly instructed to include *every* aspect they deemed relevant before rating overall quality. The choice to not having coloration rated as a separate attribute was also motivated by the observation that deficiencies in coloration are very well reflected by the ratings of overall quality as long as there are no dominant deficiencies in localization [14, 30].

All ratings were given on a unipolar scale from "schlecht" (bad) to "sehr gut" (excellent), with equi-distant points on the scale marked as "dürftig" (poor), "ordentlich" (fair), and "gut" (good). In addition to the labeled scale points, subjects could also choose one point inbetween, resulting in a nine-point scale, where "bad" corresponds to 1 and "excellent" to 9. Sometimes, bipolar scales are preferred in this context, which would give additional information about the direction of the perceived difference and in addition avoid bias for very small differences. On the other hand, the unipolar scale used here has the advantage that subjects having experience in scaling experiments are familiar with it, and that the number of attributes that can be investigated in a given time is higher compared to using a bipolar scale: for example, instead of rating agreement with respect to "source direction" on a unipolar scale, both horizontal and vertical excursions would have been needed to rate with bipolar scales.

For each attribute, nine different scenes (three loudspeaker positions, each presented three times) were presented in random order. In each scene, subjects could switch at will between the randomized BRIR sets and between loudspeaker and headphone reproduction and could listen as long as desired. Listeners rated the conditions by adjusting sliders on a graphical user interface displayed in front of them (see Fig. 1). The sliders for all six BRIR sets were displayed simultaneously, reflecting the current ratings. As proposed in [31], the order in which the sliders were displayed could be sorted according to the current ratings, and subjects could re-adjust their ratings as often as desired. Once the subject felt comfortable with the ratings given for all six BRIR sets, they moved on to the next scene.

The order in which the attributes were rated remained always the same: first all scenes were rated for "reverberance", then for "source width", and so forth in the order the attributes are given above. The advantage of the attributes not being randomized is that subjects could better focus on the respective attribute. The drawback is that learning effects could yield inconsistencies between the ratings across attributes. Since the focus of this work was on comparing different BRIR sets rather than comparing different attributes, we accepted this risk in favor of simplifying the subjects' task. As will be seen in Section 3.1, the results indicate that learning effects were indeed negligeable.

The test proceeded as follows: at the beginning of the listening test, the supervisor instructed the subject about the task and the first of the five attributes. A brief training phase (approximately 5 min) preceded the actual test, allowing the subjects to get familiar with the environment, the equipment and the experimental task. After completion of the ratings for one attribute, the following attribute was explained to the subject before he or she proceeded to give ratings concerning that attribute, without new training phases. At the beginning of each scene, the supervisor manually re-calibrated the head tracker, by asking the subject to look straight ahead for a few seconds. In total, subjects needed between 90 and 180 min, split in two sessions, to rate all scenes for all attributes.

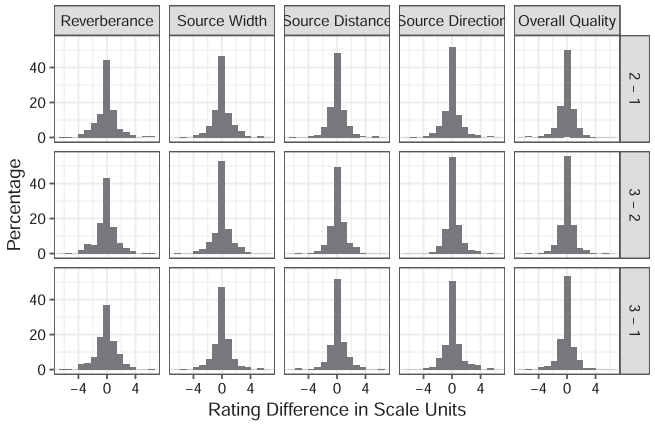

Figure 4. Histograms of rating differences between all three presentation pairs (2-1, 3-2, 3-1), for all attributes. Two scale units correspond to the difference between adjacent scale categories.

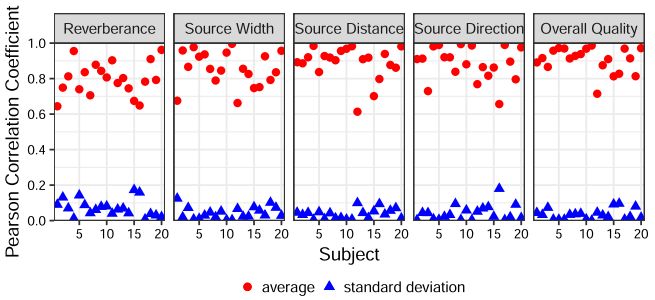

Figure 5. Consistency across the three presentations, as measured by the average and the standard deviation of Pearson correlation coefficients (over 18 combinations of BRIR set and loudspeaker position) for all three presentation pairs (2–1, 3–2, 3–1).

3 Results

3.1 Test-retest analysis

As a first analysis, the differences between ratings given in repeated presentations of the same stimulus were analyzed, see Figure 4: Despite the challenging task, the test-retest reliability was high – depending on attribute and presentation pair, between 37% and 56% of the ratings were identical across two presentations (zero difference) and between 72% and 88% were within ±1 scale units. Moreover, the distribution of the rating differences appears to be highly symmetric with respect to zero difference, and in addition very similar for all presentation pairs and attributes. As the presentation number reflects the order in which the presentations were given (first presentation 1, then 2 and then 3) and the attributes were presented in the order of the columns in Figure 4 from left to right, it can be concluded that possible temporal (e.g., learning) effects did not entail any bias in the ratings over time.

The differences in ratings given by each individual subject in repeated presentations of the same stimulus can also be used to assess the subject's ability to give consistent ratings and, eventually, to identify subjects to be discarded from the analysis because of insufficient consistency. To this end, the average and standard deviation of Pearson correlation coefficients between different presentations of the same stimulus for a number of subsequent presentations could be used as proposed in [32]. Since consistency can be assumed to not only vary between subjects but also between attributes (for some attributes it is harder to give consistent ratings than for others), these metrics were, in a first step, computed for all subjects per attribute, see Figure 5. Each single correlation coefficient referred to 18 ratings (six BRIR sets times three loudspeaker positions).

Inspection of Figure 5 suggests that higher standard deviations of the correlation coefficients often coincided with low average correlation coefficients and hence the consistency criterion could be based on the average Pearson correlation coefficient alone. This has the additional advantage that it can easily be related to Cronbach's (standardized) coefficient α,

$$\bar{r} = \frac{\alpha}{n + (1 - n)\alpha}, \tag{1}$$

as already pointed out in Cronbach's original paper [33]. Here, n is the number of presentations ($n = 3$) and $\bar{r}$ the average Pearson correlation coefficient. Often, $\alpha > 0.8$ is considered "good". For $n = 3$, this corresponds to an average Pearson correlation coefficient of $\bar{r} > 0.57$, which is surpassed by all subjects for any attribute, see Figure 5. Although it is markedly lower than the suggestion in [32], we chose this limit for an acceptable consistency, i.e., we retained all subjects in the analyses. In fact, only subtleties would change in the following analyses if more restrictive consistency criteria were applied. For the convenience of the reader, the raw data of the listening test is available at http://doi.org/10.5281/zenodo.3835811.

3.2 Ratings averaged over presentations

Given that no systematic differences between the three presentations could be spotted, the ratings were averaged over the three presentations to give, per subject, one average rating for every combination of attribute, BRIR set and loudspeaker position. These average ratings are shown in Figure 6, as a function of the BRIR set, for all combinations of attribute and loudspeaker position.

It is apparent from Figure 6 that the average ratings do not differ much across loudspeaker positions. To test this assumption, a robust repeated measures analysis of variance (rRMANOVA) on 20% trimmed means, based on the bootstrap-t-method [34] was performed for each combination of attribute and BRIR set, after verifying that normality could not be assumed for the majority of all those combinations (by visually inspecting histograms and performing Shapiro–Wilk tests).

In eight out of the 30 possible combinations, the rRMANOVA results suggested to reject the null hypothesis of equal 20% trimmed means of the ratings across loudspeaker positions ($p < 0.05$). These combinations were:

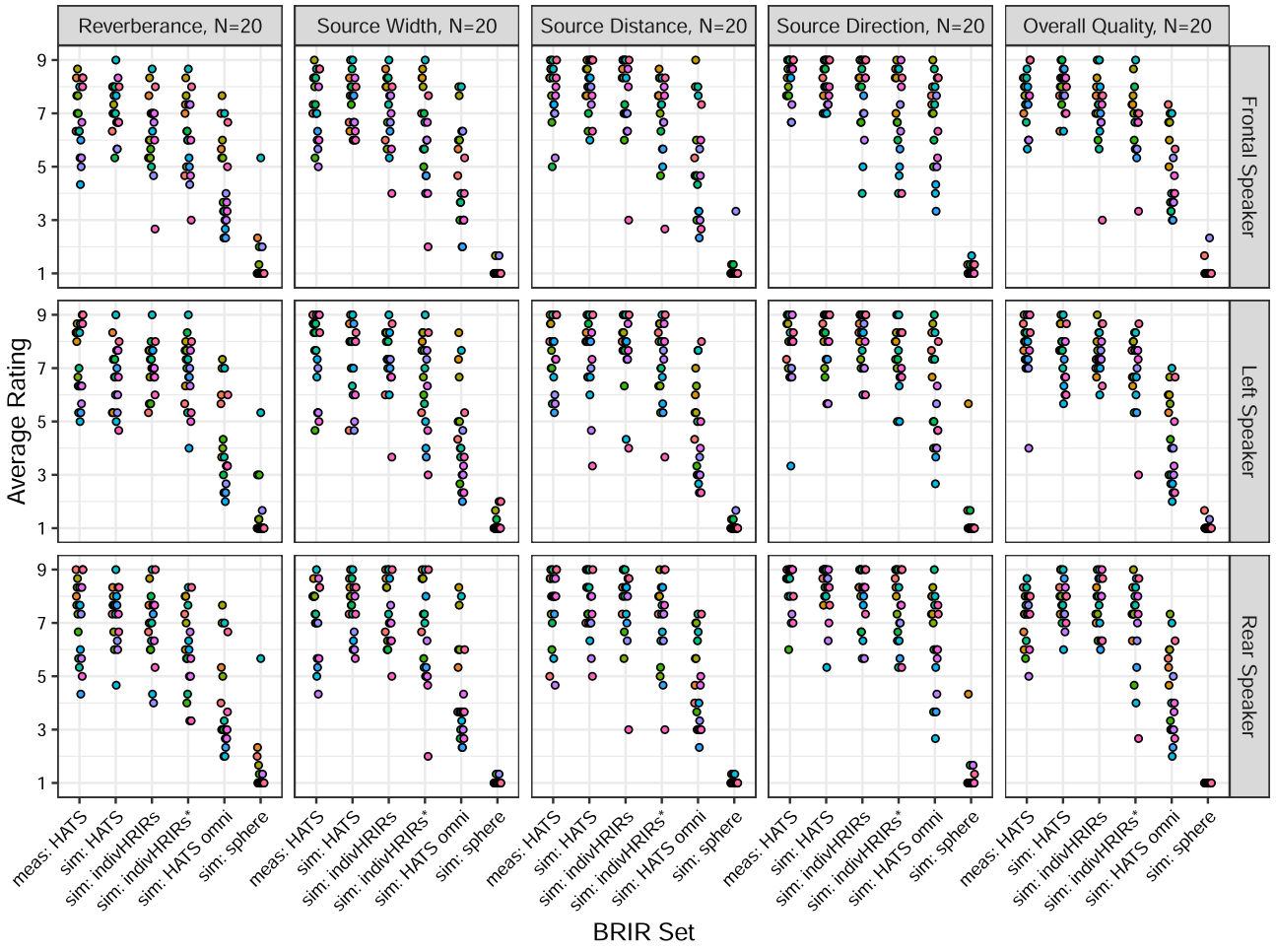

Figure 6. Individual ratings, averaged over the three presentations, as a function of the BRIR set, for all attributes and loudspeaker positions. A rating of 1 corresponds to "bad", 9 to "excellent".

"sim: HATS", "sim: indivHRIRs" and "sim: HATS omni" for "reverberance", "meas: HATS" for "source width", "sim: HATS omni" for "source direction", and "sim: indivHRIRs", "sim: indivHRIRs*" and "sim: HATS omni" for "overall quality". Since for the majority of all combinations (22 out of 30), no significant effect of loudspeaker position was observed, we refrained from further pursuing a detailed analysis for every loudspeaker position and instead proceeded with ratings that were in addition averaged over loudspeaker positions.

3.3 Ratings averaged over presentations and loudspeaker positions

In Figure 7, the ratings averaged over presentations and loudspeaker positions are shown as box plots, as a function of the BRIR set, for all attributes.

The dependency of these average ratings on the BRIR set is very similar for every attribute: the best ratings were always obtained for the first three BRIR sets (measured HATS BRIRs and simulated BRIRs using HATS and individual HRIRs), without obvious differences between them and medians between slightly below seven and and slightly above eight on the nine-point scale.

The fourth BRIR set (simulated BRIRs using own HRIRs rendered with the VAH technology) gave almost as good average ratings as the top three BRIR sets, with a slight tendency to lower ones (medians around seven).

The worst average ratings were obtained for the anchor (simulations with sphere HRIRs), essentially close to one for all attributes.

The fifth BRIR set (simulated BRIRs based on HATS HRIRs and omnidirectional sources) was rated somewhere inbetween the top group and the anchor, with medians mostly between three and five, except for "source direction" where the median was close to seven.

With the exception of "reverberance", at least two of the BRIR sets based on simulations gave median ratings of seven or above on the nine-point scale used in this study.

For "reverberance", the absolute ratings obtained with the top three BRIR sets were somewhat lower than those obtained for the other attributes.

In order to test the observed differences statistically, a rRMANOVA on 20% trimmed means was performed for

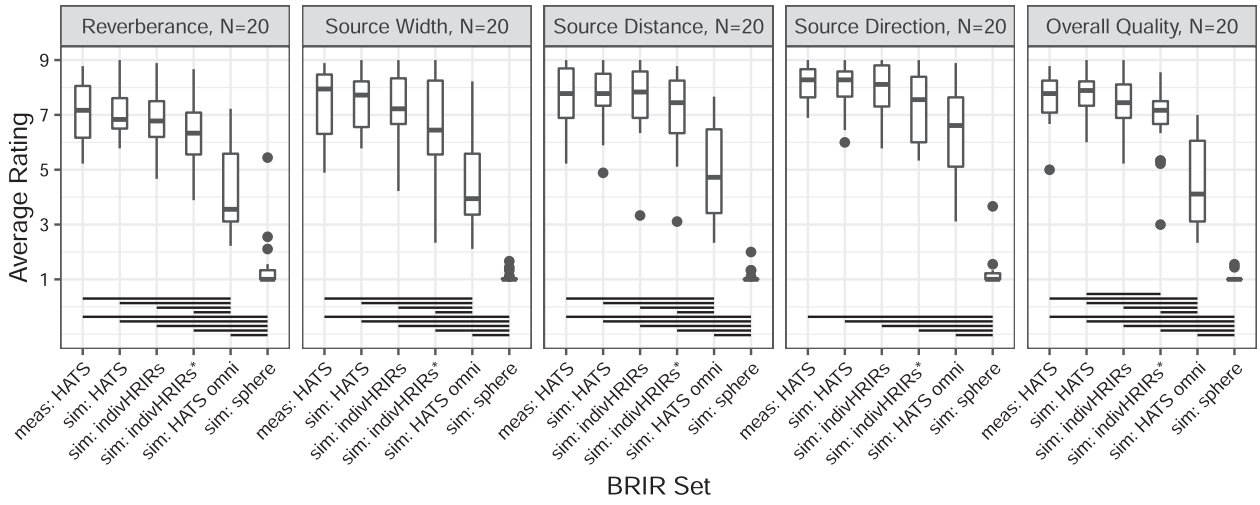

Figure 7. Box plots of ratings, averaged over the three presentations and the three loudspeaker positions, as a function of the BRIR set, for all attributes. A rating of 1 corresponds to "bad", 9 to "excellent". Statistically significant differences (*post hoc* pairwise tests for equal 20% trimmed means based on the bootstrap-t-method, $p < 0.05$) are indicated by lines in the bottom part of the facets.

each attribute. This was again motivated by observed departures from normality of the average ratings, as determined by Shapiro–Wilk tests. As may be obvious from Figure 6, the ratings for the anchor were not normally distributed, instead they were strongly skewed to the low end of the scale. But even if the anchor is omitted, there was no attribute for which the average ratings for all of the other BRIRsets could be assumed to be normally distributed. In most cases, some of them were skewed (often, but not always, to higher ratings). Note that this would also preclude the use of the otherwise popular non-parametric Wilcoxon signed rank test for *post hoc* analyses since the latter requires symmetric distributions of pairwise differences, which are rare if part of the ratings are skewed and others not.

The results of the rRMANOVA indicate that for any of the attributes, the null hypothesis of equal 20% trimmed means of the ratings for all BRIR sets should be rejected ($p < 0.05$), which is not surprising given the huge difference between any of the BRIR sets and the anchor.

To compare individual BRIR sets, *post hoc* tests for pairwise equal 20% trimmed means, again based on the bootstrap-*t*-method, see [34], were performed. These tests were chosen because in [35] it was shown for a similar setting ($N = 15$, four dependent groups to be compared, nominal simultaneous probability coverage $1 - \alpha = 0.95$) that they compared favourably against standard Bonferroni-corrected pairwise *t*-tests and non-bootstrap pairwise tests for equal 20% trimmed means. Significant differences ($p < 0.05$) are shown as lines between BRIR sets in the bottom part of Figure 7.

The results are essentially similar for all of the attributes, with a slightly different variant for "source direction": First, for all attributes, the anchor was rated significantly worse than any other BRIR set. Second, the fifth BRIR set (simulated BRIRs based on HATS HRIRs and omnidirectional sources) was rated significantly worse than any of the first four BRIR sets, for all attributes except "source direction". For "source direction", the ratings for the fifth BRIR set were substantially improved in comparison to the other attributes, such that the statistical significance of the differences to the first four BRIR sets vanished. The observed slight degradations associated with the VAH technology (fourth BRIR set compared to the top three ones) were mostly non-significant, with the exception of "overall quality" where the difference to simulations based on HATS HRIRs (second BRIR set) was significant.

4 Discussion

4.1 Agreement between auralization and reality

For all attributes except "reverberance", there were BRIR sets for which the agreement with the real room was rated with medians of 7.5 or better on the nine-point scale used. Since subjects hesitate to use the very extreme of the scale, in particular if the reference is not hidden, we consider this "close to reality".

Interestingly, both measured and simulated BRIR sets were capable of eliciting such high ratings, indicating that realistic auralizations of speech are possible for simulated rooms as well. This highlights the progress that has been made in the simulations using the RAZR package since the round robin on auralization [10]. There, the omnidirectional version of RAZR was also used (algorithm B) and performed well in the perceptual evaluation, whereas in the present study, that version was clearly superseeded by all other variants investigated except the anchor.

For "reverberance", the ratings for the best BRIR sets were slightly lower than for the other attributes, with medians around seven on the nine-point scale. Again, this was true for both measured and simulated BRIR sets.

Since the measured BRIR set gave similar average ratings as the best simulated BRIR sets, the slight decrease of ratings compared to the other attributes cannot be attributed to inaccurracies of the room simulation alone. It should be noted that "reverberance" was reported the hardest to rate by the subjects, which is also reflected in partly mediocre consistency metrics (see Fig. 5). Also, "reverberance" was the first attribute to be judged in the listening test. Hence it must be suspected that the slightly lower ratings may rather reflect the uncertainty of the subjects in judging this attribute, resulting in stronger hesitation to give high ratings.

4.2 Simulated versus measured BRIRs

No significant differences between ratings given for measured BRIRs and those given for the best simulated BRIRs were observed in the current study.

This is in line with the results from [8] where it was shown that the difference between ratings of room-acoustical attributes (similar ones as in the present study) for measured and simulated BRIRs was within repeatability limits for most attributes. On the other hand, the general conclusion of the round robin on room acoustical simulation and auralization [10] was that in contrast to measurement-based auralizations, simulation-based auralizations were not yet capable of accurately predicting perceptual properties of sound sources in virtual acoustic environments. How can these apparently conflicting conclusions be explained?

First, in both [8] and the present study, important simulation parameters were adjusted such that certain metrics from simulation and measurements agreed within reasonable limits: in [8], absorption and scattering coefficients were adjusted such that reverberation and clarity metrics were met, whereas in the present study, only absorption coefficients were adjusted. In both studies, these adjustments referred to a few selected sender–receiver pairs and to impulse responses for omnidirectional receivers, i.e., *not* to the BRIRs used in the perceptual evaluation. Still, the simulations we used incorporated essential knowledge obtained from the measurements, which will have contributed to them sounding more realistic. Adjusting simulation parameters to measurements is of course not possible if rooms are to be auralized that do not yet exist. However, besides serving as a benchmark, many applications can be thought of where an existing room is to be auralized and such adjustments could be made.

Second, whereas in [8] and in the present study speech was used as primary stimulus (plus singing in [8]), both chamber music and pulsed pink noise were used in [10]. Since it is well-known that speech is the least critical stimulus suitable for detecting small deficiencies of virtual renderings, it is not surprising that [10] found larger differences between measurement and simulation. Yet speech remains an important stimulus with regard to possible applications.

Third, the conclusion in [10] was based on a statistical analysis of *all* simulation packages and *all* stimuli. If instead the best simulation package and music stimuli had been considered only, the rated differences between simulation and measurement would have been observed to vanish almost completely, see Figures 9 and 10 in [10].

4.3 Do we need individual HRIRs?

As already noted in [2], the use of individual HRIRs in the simulations does not appear to result in any benefit over using generic (HATS) HRIRs.

This is surprising at first sight, because non-individual HRIRs are often associated with an impaired perception of spatial sound, resulting in front-back- or up-down confusions [36, 37] and lacking externalization ("in-head-localization" [38]).

On the other hand, as early as in 1976, it was shown that out-of-head localization can be promoted by adding room reflections to binaural stimuli [39]. More recently, it was observed that room reflections improve externalization of speech signals [40, 41]. In [1], it was found that subjects were less able to detect differences between real and auralized stimuli with increasing reverberation and decreasing direct-to-reverberant energy ratios (DRRs), albeit using individual BRIRs. This effect was particularly pronounced for speech stimuli. They argued that in increasingly reflective environments, the "spectral differences between reality and simulation are partially smoothed out", i.e. small deficiencies will be masked. Although not explicitly considered in that study, one could use the same argument (masking small differences by reflections) and apply it to differences between individual and HATS HRIRs.

Also, it has been shown that by using dynamic (i.e., head-tracked) binaural presentation instead of static presentation, the externalization of speech [42] and pulsed noise [43] stimuli is improved. In [43] it was in addition found that individual HRIRs resulted in better localization of pulsed noise stimuli *only* in static presentation scenarios (if at all) whereas no significant differences between individual and HATS HRIRs were found for the head-tracked presentations.

This raises the question whether individual HRIRs are needed for the auralization of speech at all? An important aspect seems to be that speech is a stimulus that is known to be not very critical. Also, we must recall that, by performing the experiments in the real room in the present study, all other cues including visual ones were present naturally, thereby providing the best possible conditions for a realistic auralization. It is likely that lacking externalization and front-back-confusions become more relevant if auditory-only cues are presented. In addition, our results strictly hold only for the room studied here and for head-tracked binaural auralizations. If reverberation is decreased (e.g, in very dry recording rooms or in outdoor scenarios), or if auralizations are limited to static binaural renderings, the importance of individual HRIRs/BRIRs may be more prominent. It is also conceivable that there might be attributes which were not tested in the present study but for which a larger effect of individual over HATS HRIRs would be obtained.

Still, there is hope that for many applications, non-individual HRIRs may be sufficient even if the auralization

takes place in a different room – possibly by augmenting the auralization with visualization.

4.4 Effect of ignoring source directivity

The critical progress over the room simulations used in [2] was that, in the present study, the source directivities could be properly taken into account. In order to investigate the effect of ignoring source directivities, one BRIR set was included in the present study in which omnidirectional sources were assumed in the simulation. As shown in Section 3.3, this BRIR set was rated significantly worse than the BRIR sets in which the correct source directivities were used in the simulation, for all attributes except "source direction". Since "source direction" depends the least on a correct room simulation (because it is dominanted by the direct sound), this suggests that the lower ratings are due to incorrect room simulations.

It may seem surprising though to see that the ratings for this BRIR set were *this* low in the present study. In [2], a similar BRIR set (called "HATS sim" there) was rated much better. More precisely, medians of the ratings were about two scale points higher, except for "source distance" (1.4 points higher) and "source direction" (0.6 points higher). It must be noted, however, that the adjustment of the absorption coefficients was based on simulations with omnidirectional sources in [2] and on simulations with directional sources in the present study. This again will affect attributes such as "reverberance" much more than "source direction". Hence, the discrepancy between the ratings for BRIR sets with ignored source directivities in the two studies can at least partly be explained by the mismatch of absorption coefficients.

But even if the higher ratings from [2], where absorption coefficients were adjusted for the simulation with omnidirectional sources, are taken for a comparison, the ratings obtained in the present study for the simulated HATS (including proper source directivities) are almost two scale points higher, confirming the importance of correctly mimicking the source directivities in the simulations.

4.5 Implications for the virtual artificial head (VAH) technology

Parallel to the discussion on the differences between auralized and real rooms, the current study offered a new way to assess the VAH technology in a more relevant acoustical environment than before – previous studies, e.g. [14, 44], were limited to static free-field scenarios only.

Depending on the attribute, median ratings for the BRIR sets created with the VAH technology were close to "good" (seven out of nine on the scale used). They were always slightly lower (typically about one half scale point) than median ratings for the corresponding original BRIR sets, but with one exception not significantly so. This is encouraging while also demanding further improvement.

In VAH design, many trade-offs have to be dealt with, regarding the number of microphones, array extension and topology, and optimization parameters. If the number of microphones avaiable is limited to moderate values (say, 32), then it will be impossible to guarantee accuracy and robustness for all possible directions across the whole frequency range of interest. Hence, one has to prioritize certain directions over others. For the current application (speech in lecture room), it appeared to be a good strategy to set constraints on horizontal directions only, although room reflections occurred from many different directions. This is, however, just a starting point, and other optimization targets may be thought of in future work.

4.6 Role of the anchor

The results of the listening test clearly show that the anchor was always rated distinctly worse than any other BRIR set, which is not surprising given the large spectral deviation to the other stimuli, shown in Figure 3. It may, however, raise concerns that the low ratings for the anchor could explain the high ratings given to most other conditions.

It should be noted, however, that the subjects were presented with explicit labels on the scale they used that indicated the perceived quality, and that the comparison between perceptually similar BRIR sets was facilitated by having the subjects deliberately switch between BRIR sets and arrange the order the BRIR sets according to their current rating, see Section 2.3.

One can also draw a rough comparison to the round robin on auralization [10]. As already mentioned above, the omnidirectional version of RAZR was also used there (algorithm B), albeit in a slightly different test design (comparison between auralized and measured BRIRs instead of comparison between auralized BRIRs and the real scenario in the present study). The configurations in the round robin that are comparable to our test are the ones that use a unipolar scale and a similar stimulus, i.e., "difference" and "source position", both for music stimuli and the small room.

First, for "difference", a median scale value of slightly below 0.5 was obtained for music stimuli in the small room in the round robin [10]. This attribute is comparable to our difference in "overall quality". The median scale value found for the fifth BRIR set (HATS auralization using the omnidirectional version of RAZR) is slightly above four on our nine-point scale, see Figure 7, which would correspond to about 0.4 on a 0…1 scale, i.e. even worse compared to the result of the round robin.

Second, for "source position", a median scale value of 0.7 was found in the round robin [10], whereas the corresponding result in the present study ("source direction", fifth BRIR set) is 6.6 on the nine-point scale (corresponding to 0.7 on a 0…1 scale), see Figure 7.

Both comparisons suggest that comparable results were obtained in both studies and consequently, it seems to be unlikely that the ratings for the "better" BRIR sets were markedly influenced by the anchor.

It should finally be noted that substantially improved versions of analytical BRIR sets can be obtained, e.g. by a better choice of loudspeaker and/or headphone

compensation for this specific BRIR set, but this was not a primary concern in the current study.

4.7 Perception transfer to the reproduction room

As stated above, the experiments reported here took place in the real room and hence, except for auditory cues during headphone renderings, all cues were present naturally.

In many applications, auralizations will be listened to in a different reproduction room (typically a laboratory), for instance in cases where the real room does not (yet) exist or if access to the real room is restricted. In such cases, the goal is to transfer the perceptions that would have been created in the real room to the virtual one, presented in the reproduction room. This is known to be a challenging task, since for instance externalization of virtual sources may break down to a large degree if the virtual room differs acoustically from the reproduction room [45].

If auralizations are to be used as a tool to base decisions on in such cases, it will be crucial that all relevant perceptual attributes be preserved in the auralization. Since a direct comparison to a reference in perceptual tests is only possible either in the real room *or* in the reproduction room, ABX tests alone will not be sufficient to ensure the perception transfer from the real to the reproduction room and will therefore have to be at least complemented by scaling experiments.

Additional visual stimuli may be helpful, but there is clearly more research needed to better understand the opportunities, challenges, and limits of these techniques.

5 Conclusion and outlook

From the results of the present study, one can conclude that close-to-real binaural auralizations of speech are possible if all modalities (auditory, visual, ...) are properly reproduced. This is not only true for auralizations based on measured BRIRs but also for auralizations based on simulated BRIRs, if the simulations are tuned to closely match the reverberation time of the real room. The requirement on the reproduction accuracy for auditory cues can be relaxed in that the use of generic HRIRs is sufficient for auralizations of speech in reverberant spaces, provided all other modalities are correctly reproduced.

Further work is required to clarify the validity of these findings for acoustically dryer environments and to investigate under which conditions realistic auralizations can be obtained if they cannot be presented in the real room but have to be transferred to a different listening environment. The use of head-mounted displays to augment the auralization by providing associated visual cues seems to be promising in this respect.

Acknowledgments

We would like to thank the two anonymous reviewers for their thoughtful comments on an earlier version of the manuscript. This work was partly funded by Bundesministerium für Bildung und Forschung under grant no. 03FH021IX5, and by Deutsche Forschungsgemeinschaft (DFG, German research Foundation) – project IDs 352015383 (SFB 1330 C5) and 444832396 (SPP 2236). We also want to thank our subjects for their participation in the study.

Conflict of interest

Author declared no conflict of interests.

References

1. F. Brinkmann, A. Lindau, S. Weinzierl: On the authenticity of individual dynamic binaural synthesis. The Journal of the Acoustical Society of America 142 (2017) 1784–1795.
2. M. Blau, A. Budnik, S. van de Par: Assessment of perceptual attributes of classroom acoustics: Real versus simulated room. Proceedings of the Institute of Acoustics 40 (2018) 556–564. http://www.proceedings.com/42138.html.
3. M. Kleiner, B.-I. Dalenbäck, P. Svensson: Auralization-an overview. JAES 41 (1993) 861–875.
4. D. Thery, V. Boccara, B.F.G. Katz: Auralization uses in acoustical design: A survey study of acoustical consultants. The Journal of the Acoustical Society of America 145 (2019) 3446–3456.
5. M. Vorländer: Auralization – Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality, RWTHedition. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
6. T. Lokki, V. Pulkki: Evaluation of geometry-based parametric auralization, in Proceedings of the AES 22nd International Conference: Virtual, Synthetic, and Entertainment Audio, Audio Engineering Society. 2002.
7. J.H. Rindel, C. Lynge Christensen: Room acoustic simulation and auralization – How close can we get to the real room? in Proc. 8th Western Pacific Acoustics Conference, Melbourne (Australia). 2003, 1025J p.
8. B.N.J. Postma, B.F.G. Katz: Perceptive and objective evaluation of calibrated room acoustic simulation auralizations. The Journal of the Acoustical Society of America 140 (2016) 4326–4337.
9. T. Wendt: A Computationally Efficient and Perceptually Plausible Room Acoustics Simulation Method. Dissertation. Universität Oldenburg, 2018.
10. F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, S. Weinzierl: A round robin on room acoustical simulation and auralization. The Journal of the Acoustical Society of America 145 (2019) 2746–2760.
11. K.-H. Brandenburg, S. Werner, F. Klein, C. Sladeczek: Auditory illusion through headphones: History, challenges and new solutions, in Proc. 22nd Int. Congr. on Acoustics, Buenos Aires (Argentina). 2016.
12. J. Oberem, B. Masiero, J. Fels: Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods. Applied Acoustics 114 (2016) 71–78.
13. T. Wendt, S. van de Par, S.D. Ewert: A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation. JAES 62 (2014) 748–766.

14. E. Rasumow, M. Blau, S. Doclo, S. van de Par, M. Hansen, D. Püschel, V. Mellert: Perceptual evaluation of individualized binaural reproduction using a virtual artificial head. The Journal of the Audio Engineering Society 65 (2017) 448–459.
15. M. Fallahi, M. Hansen, S. Doclo, S. van de Par, D. Püschel, M. Blau: Individual binaural reproduction with high spatial resolution using a virtual artificial head with a moderate number of microphones. The Journal of the Acoustical Society of America (submitted) (2020).
16. M.C. Vigeant, L.M. Wang, J.H. Rindel: Investigations of multi-channel auralization technique for solo instruments and orchestra, in Proc. 19th Int. Congr. on Acoustics, Madrid (Spain). 2007.
17. B.N.J. Postma, H. Demontis, B.F.G. Katz: Subjective evaluation of dynamic voice directivity for auralizations. Acta Acustica United With Acustica 103 (2017) 181–184.
18. H. Steffens, S. van de Par, S.D. Ewert: Perceptual Relevance of Speaker Directivity Modelling in Virtual Rooms, in Proc. 23rd Int. Congr. on Acoustics, Aachen (Germany). 2019, pp. 2651–2658.
19. J. Poppitz, M. Blau, M. Hansen: Entwicklung und Evaluation eines Systems zur Messung individueller HRTFs in privater Wohn-Umgebung, in Fortschritte der Akustik – DAGA 2016, Aachen. 2016.
20. P. Majdak, P. Balazs, B. Laback: Multiple exponential sweep method for fast measurement of head-related transfer functions. The Journal of the Audio Engineering Society 55 (2007) 623–637.
21. A. Novák, L. Simon, F. Kadlec, P. Lotton: Nonlinear system identification using exponential swept-sine signal. IEEE Transactions on Instrumentation and Measurement 59 (2010) 2220–2229.
22. M.S. Puckette: Pure data: Another integrated computer music environment, in Proceedings of the International Computer Music Conference, San Francisco (USA). 1996.
23. E. Rasumow, M. Blau, M. Hansen, S. van de Par, S. Doclo, V. Mellert, D. Püschel: Smoothing individual head-related transfer functions in the frequency and spatial domains. The Journal of the Acoustical Society of America 135 (2014) 2012–2025.
24. E. Rasumow, M. Hansen, S.v.d. Par, D. Püschel, V. Mellert, S. Doclo, M. Blau: Regularization approaches for synthesizing HRTF directivity patterns. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (2016) 215–225.
25. R.O. Duda, W.L. Martens: Range dependence of the response of a spherical head model. The Journal of the Acoustical Society of America 104 (1998) 3048–3058.
26. O. Kirkeby, P.A. Nelson: Digital filter design for inversion problems in sound reproduction. The Journal of the Audio Engineering Society 47 (1999) 583–595.
27. P.D. Hatziantoniou, J.N. Mourjopoulos: Generalized Fractional-Octave Smoothing of Audio and Acoustic Responses. The Journal of the Audio Engineering Society 48 (2000) 259–280.
28. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet, Cambridge University Press, 1999.
29. H. Jaeger, J. Bitzer, U. Simmer, M. Blau: Echtzeitfähiges binaurales Rendering mit Bewegungssensoren von 3D-Brillen, in Fortschritte der Akustik – DAGA 2017, Kiel. 2017.
30. M. Fallahi, M. Blau, M. Hansen, S. Doclo, S. van de Par, D. Püschel: Optimizing the microphone array size for a virtual artificial head, in Proc. International Symposium on Auditory and Audiological Research (ISAAR) 2017. 2017.
31. P. Chevret, E. Parizet: An efficient alternative to the paired comparison method for the subjective evaluation of a large set of sounds, in Proc. 19th Int. Congr. on Acoustics, Madrid (Spain). 2007.
32. A. Andreopoulou, B. Katz: Investigation on subjective HRTF rating repeatability, in Proc. 140th AES Convention 2016, Paris (France), Audio Engineering Society. 2016.
33. L.J. Cronbach: Coefficient alpha and the internal structure of tests. Psychometrika 16 (1951) 297–334.
34. R.R. Wilcox: Introduction to Robust Estimation and Hypothesis Testing, 4th ed. Elsevier, 2017.
35. R.R. Wilcox: Pairwise comparisons using trimmed means or M-estimators when working with dependent groups. Biometrical Journal 39 (1997) 677–688.
36. P. Damaske, B. Wagener: Richtungshörversuche über einen nachgebildeten Kopf. Acustica 21 (1969) 30–35.
37. E.M. Wenzel, M. Arruda, D.J. Kistler, F.L. Wightman: Localization using nonindividualized head-related transfer functions. The Journal of the Acoustical Society of America 94 (1993) 111–123.
38. G. Plenge: Über das Problem der Im-Kopf-Lokalisation. Acustica 26 (1972) 241–252.
39. N. Sakamoto, T. Gotoh, Y. Kimura: On "Out-of-Head Localization" in headphone listening. The Journal of the Audio Engineering Society 24 (1976) 710–716.
40. J. Catic, S. Santurette, T. Dau: The role of reverberation-related binaural cues in the externalization of speech. The Journal of the Acoustical Society of America 138 (2015) 1154–1167.
41. H.G. Hassager, F. Gran, T. Dau: The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment. The Journal of the Acoustical Society of America 139 (2016) 2992–3000.
42. W.O. Brimijoin, A.W. Boyd, M.A. Akeroyd: The contribution of head movement to the externalization and internalization of sounds. PLoS One 8 (2013) e83068.
43. J. Oberem, J.-G. Richter, D. Setzer, J. Seibold, I. Koch, J. Fels: Experiments on localization accuracy with non-individual and individual HRTFs comparing static and dynamic reproduction methods. bioRxiv (2020). https://doi.org/10.1101/2020.03.31.011650.
44. M. Fallahi, M. Hansen, S. Doclo, S. van de Par, D. Püschel, M. Blau: Individual binaural reproduction of music recordings using a virtual artificial head, in AES Conference e-Brief 61. 2018.
45. S. Werner, F. Klein, T. Mayenfels, K. Brandenburg: A summary on acoustic room divergence and its effect on externalization of auditory events, in 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX). 2016, pp. 1–6.

Cite this article as: Blau M, Budnik A, Fallahi M, Steffens H, Ewert S, et al. 2021. Toward realistic binaural auralizations – perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario. Acta Acustica, **5**, 8.