ASSESSING THE AUTHENTICITY OF INDIVIDUAL DYNAMIC BINAURAL SYNTHESIS

Fabian Brinkmann, Alexander Lindau, Martina Vrhovnik, Stefan Weinzierl

Audio communication group,
Technical University of Berlin, TUB
Berlin, Germany
fabian.brinkmann@tu-berlin.de

ABSTRACT

Binaural technology allows to capture sound fields by recording the sound pressure arriving at the listener's ear canal entrances. If these signals are reconstructed for the same listener the simulation should be indistinguishable from the corresponding real sound field. A simulation fulfilling this premise could be termed as perceptually *authentic*.

Authenticity has been assessed previously for static binaural resynthesis of sound sources in anechoic environments, i.e. for HRTF-based simulations not accounting for head movements of the listeners. Results indicated that simulations were still discernable from real sound fields, at least, if critical audio material was used.

However, for *dynamic* binaural synthesis to our knowledge – and probably because this technology is even more demanding – no such study has been conducted so far. Thus, having developed a state-of-the-art system for individual dynamic auralization of anechoic and reverberant acoustical environments, we assessed its perceptual authenticity by letting subjects directly compare binaural simulations and real sound fields. To this end, individual binaural room impulses were acquired for two different source positions in a medium-sized recording studio, as well as individual headphone transfer functions. Listening tests were conducted for two different audio contents applying a most sensitive ABX test paradigm. Results showed that for speech signals many of the subjects failed to reliably detect the simulation. For pink noise pulses, however, all subjects could distinguish the simulation from reality. Results further provided evidence for future improvements.

1. INTRODUCTION

As overall criteria for the quality of virtual acoustic environments, the perceived *plausibility* and *authenticity* has been proposed [1], [2]. Whereas the plausibility of a simulation refers to the degree of agreement with the listener's expectation towards a corresponding real event (agreement with an inner reference), authenticity refers to the perceptual identity with an explicitly presented real event (agreement with an external reference). While a non-individual data-based dynamic binaural synthesis has already been shown to provide plausible simulations [3], a dynamic synthesis based on individual binaural recordings appears to be a particularly promising candidate for a perceptually authentic acoustical simulation. Further, a formal assessment of the authenticity of state-of-the-art binaural technology would be of great practical relevance: Since nearly all currently known approaches to sound field synthesis (such as wave field synthesis, or higher order ambisonics) can be transcoded into binaural signals, a perceptually authentic binaural reproduction would provide a convenient reference simulation required for the strict, reliable and comprehensive evaluation of a wide variety of simulation approaches and systems [4].

Three empirical studies were found to be concerned with the authenticity of binaural simulations. However, all three studies assessed static auralization, i.e., simulations not accounting for natural head movements of the listeners. In order to allow for a convenient comparability, statistical significance of the observed results was assessed based on exact Bernoulli test statistics, if not initially given.

Langendijk and Bronkhorst [5] assessed the authenticity of individual binaural reproduction for six sound sources distributed evenly around the listener. Binaural signals were reproduced utilizing small earphones placed 1 cm in front of the concha with only little influence on the sound field of external sources. Band limited white noise bursts (500 Hz–16 kHz) were presented in a four interval 2AFC (alternative forced choice) paradigm where each sequence of four noise bursts contained three identical and one 'oddball'-stimulus in either second or third position, that had to be detected by the subjects. Detection rates across subjects were slightly but significantly above chance ($p_{\text{correct}} = 0.53$, 6 subjects, $N_{\text{total}} = 1800$ trials).

Moore et al. [6] conducted a similar listening test. Subjects participated twice in the experiment, and were considered untrained in the first run and trained in the second. A frontal sound source was auralized using cross-talk canceled (transaural) reproduction of individual binaural recordings. When presenting click or noise stimuli to trained subjects detection rates were again slightly but significantly above chance ($p_{\text{corr. click}} = p_{\text{corr. noise}} = 0.594$, 8 subjects, $N_{\text{total}} = 192$). *Untrained subjects*, however, were not able to detect the binaural simulation reliably ($p_{\text{corr. click}} = 0.5$, $p_{\text{corr. noise}} = 0.54$, $p_{\text{corr.testable}} = 0.675$ @ $\alpha = 0.05$ with 95% power, Dunn-Sidák corrected for multiple testing). Moreover, when using a *synthetic vowel sound*, the simulation was indistinguishable for both trained and untrained subjects ($p_{\text{corr.observed}} = 0.48$, $p_{\text{corr.testable}}$ as mentioned above).

Masiero [7] tested authenticity in a 3AFC test paradigm utilizing 24 sound sources distributed evenly around the listeners. Individual binaural signals were presented to 40 subjects through circumaural open headphones using noise, speech and music stimuli. Average detection rates were $p_{\text{corr. noise}} = 0.87$, $p_{\text{corr. speech}} = 0.74$, and $p_{\text{corr.music}} = 0.71$ (transformed to 2AFC detection rates for better comparability). While not being given originally by the authors, a *post hoc* inferential statistics analysis of the raw data revealed that for all three stimulus conditions detections rates were significantly above chance. Further, an

ANOVA conducted by Masiero showed the stimulus effect to be significant.

All three studies used some kind of head rest to control the subjects' head position. In addition, Moore et al. and Masiero monitored the subjects' head position with optic or magnetic tracking systems. Throughout his study, Masiero allowed for head movements between ±1°–2° rotation, and ±1–2 cm translation, respectively. Additionally, Masiero allowed his subjects to listen three times to the sequence of test stimuli whereas in the other two studies each condition was presented only once.

While – technically – being a far more demanding reproduction mode than static auralization, perceptual authenticity of *dynamic* binaural synthesis has not been assessed before. Moreover, a success of such an assessment has become more likely as number of technical improvements has been introduced recently: For example, new extraaural binaural headphones were presented *(BKsystem,* [8]) along with a perceptually optimized approach to the compensation of the headphone transfer function [9]. Further, an in-ear measurement systems for the reliable acquisition of individual binaural transfer functions (*PRECISE*, [9]) has been developed, and crossfade artifacts of dynamic binaural rendering have been minimized [10].

Further, as shown above, former studies achieved high statistical test power by cumulating test results over individuals and repeated trials while omitting a priori discussions of practical effect size and required test power. However, in order to limit the required methodological effort, and as individual performance was expected to be potentially quite different, we aimed at designing our test to produce practically meaningful results already on the level of individual subjects (cf. section 2.5).

2. METHOD

2.1. Setup

The listening tests were conducted in the recording studio of the *State Institute for Music Research*[1], Berlin ($V = 122\ \mathrm{m}^3$, $RT_{1\mathrm{kHz}} = 0.65$ s). Subjects were seated on a customized chair with an adjustable neck rest and a small table providing an armrest and space for placing the tactile interface used throughout the test (*Korg nanoKONTROL* Midi-Interface). An LCD screen was used as visual interface and placed 2 m in front of the subjects at eye level.

Two active near-field monitors (*Genelec 8030a*) were placed in front and to the right of the subjects at a distance of 3 m and a height of 1.56 m, corresponding to source positions of approximately 0° azimuth, 8° elevation (source 1) and -90° azimuth, 8° elevation (source 2). With a critical distance of 0.8 m and a loudspeaker directivity index of ca. 5 dB at 1 kHz, the source-receiver distance results in a slightly emphasized diffuse field component of the sound field. The height was adjusted so that the direct sound path from source 1 to the listening position was not blocked by the LCD screen. The source positions were chosen to represent conditions with minimal and maximal interaural time and level difference at a neutral head orientation (see test setup, Fig. 1).

Figure 1: *Listening test environment and used setup.*

For binaural reproduction, low-noise DSP-driven amplifiers and extraaural headphones were used, which were designed to exhibit only minimal influence on sound fields arriving from external sources while providing full audio bandwidth (*BKsystem*, [8]). Headphones were worn during the entire listening test, i.e. also during the binaural measurements, this way allowing for instantaneous switching between binaural simulation and corresponding real sound field. The subjects' head position was controlled using head tracking with 6 degrees of freedom (x, y, z, azimuth [head-above-torso orientation], elevation, lateral flexion) with a precision of 0.001 cm and 0.003°, respectively (*Polhemus Patriot*). A long term test of eight hours 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 earmolds. The molds were available in three different sizes, providing a good fit and reliable positioning for a wide range of individuals [9]. Phase differences between left and right ear microphones did not exceed ±2° avoiding audible interaural phase distortion [11].

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

2.2. Reproduction of Binaural Signals

The presence of headphones influences the sound field at the listeners' ears. Having considered an additional filter for compensating this effect [12], Moore et al. [6] concluded that headphones should not be used for direct comparisons of simulation and reality and consequently used transaural sound reproduction for their listening tests on authenticity. In contrast, we argue that a test on authenticity is not compromised as long as wearing the headphones (a) would affect real sound field and simulation in an identical manner and (b) would not mask possible cues for discriminating between the two. Condition (a) will be fulfilled by wearing the headphones both during measurement and simulation. For assessing condition (b), binaural room impulse responses (BRIRs) were measured with and without three types of headphones (*BKsystem*, *STAX SRS 2050 II*, *AKG K-601*) using a source at 2 m distance, -45° azimuth and 0° elevation for head-above-torso orientations in the range of ±80° azimuth. For this purpose, the head and torso simulator FABIAN equipped with a computer controlled neck joint for high precision and automated control of the head-above-torso orientation was used [13]. The headphone's influence was analyzed based on differences in the magnitude responses, and with respect to deviations of interaural time and level differences (ITD, ILD). For the *BKsystem*, magni-

[1] Staatliches Institut für Musikforschung, http://www.sim.spk-berlin.de/

tude response differences (Fig. 2, top left) show an irregular pattern with differences between approx. ±7.5 dB.

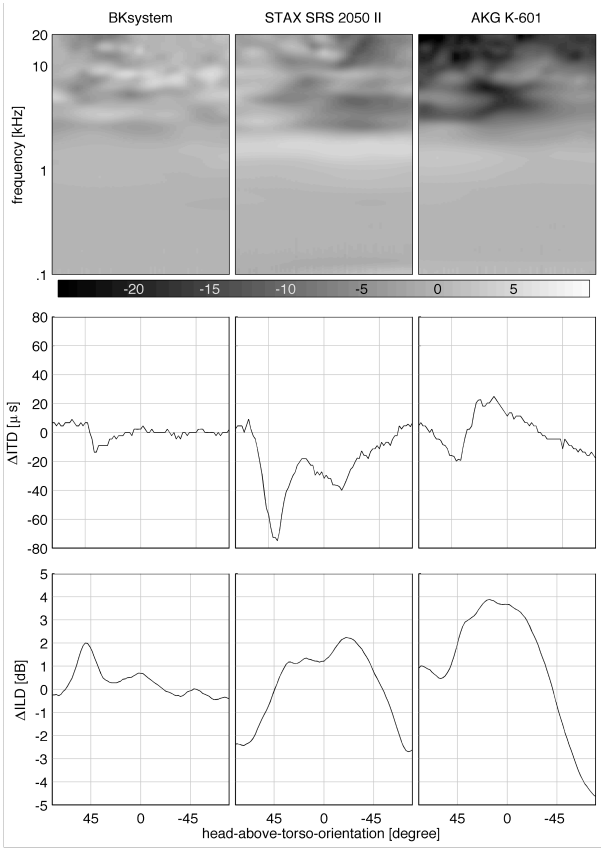

Figure 2: *Differences observed in BRIRs when measured with and without headphones for head-above-torso-orientations of between ±80° and for a source at -45° azimuth and 0° elevation. Top: Magnitude spectra (3rd octave smoothed, right ear, gray scale indicates difference in dB); Middle: ITDs; Bottom: ILDs.*

Whereas differences in magnitudes might influence localization in the median plane [14] the perceivable bandwidth of the signal remains largely unaffected making it unlikely that potential cues for a direct comparison would be eliminated. ITD and ILD differences are displayed in Fig. 2 (middle and bottom) and are believed to be inaudible for most head orientations. Assuming just audible differences of approximately 10-20 µs and 1 dB, respectively [1], only at 45°, where the ipsilateral ear is fully shadowed by the headphone, ILD differences slightly exceed the assumed threshold of audibility.

The observed differences are comparable to those found by Langendijk and Bronkhorst [5] who used small earphones near to the concha. Additionally, it is worth noting that differences were more than twice as high if conventional headphones were used (see Fig. 2).

2.3. Measurement of Individual Binaural Transfer Functions

Binaural room impulse responses and headphone transfer functions (HpTFs) were measured and processed for every subject prior to the listening test. *Matlab*® was used for audio playback, recording and processing the input signals. The head position of the subject was monitored using *Pure Data*. Communication between the programs was done by UDP messages. All audio processing was conducted at a sampling rate of 44.1 kHz.

Before starting the measurements, subjects put on the headphones and were familiarized with the procedure. Their current head position, given by azimuth and x/y/z coordinates was displayed on the LCD screen along with the target position given only by azimuth. Additionally, an acoustic guidance signal was played back through the headphones helping subjects finding the target azimuth for the subsequent measurement. The head tracker was calibrated with the test subject looking at a frontal reference position marked on the LCD screen. Subjects were instructed to keep their eye level aligned to the reference position during measurement and listening test, this way establishing also indirect control over their head elevation and roll. For training proper head-positioning, subjects were instructed to move their head to a specific azimuth and hold the position for 10 seconds. All subjects were quickly able to maintain a position with a precision of ±0.2° azimuth.

Then, subjects inserted the measurement microphones into their ear canals until they were flush with the bottom of the concha. Correct fit was inspected by the investigator. The measurement level was adjusted to be comfortable for the subjects while also avoiding limiting of both the DSP-driven loudspeakers and headphones.

BRIRs were measured for head-above-torso orientations between ±34° in azimuth and with a resolution of 2° providing smooth adaption to head movements [15]. The range was restricted to allow for a comfortable range of movements and convenient viewing of the LCD screen. Sine sweeps of an FFT order 18 were used for measuring transfer functions achieving a peak-to-tail signal-to-noise ratio (SNR) of approx. 80 dB for the BRIR at neutral head orientation without averaging [16].

The subjects started a measurement by pressing a button on the MIDI-interface after moving their head to the target position and reached it within ±0.1°. For the frontal head orientation, the target orientation had to be met also 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 subjects were instructed to meet the targeted azimuth only. During the measurement, head movements of more than 0.5° or 1 cm would have led to a repetition of the measurement, which rarely happened. These tolerance levels were set in order to avoid audible artifacts introduced by imperfect positioning [1][17].

Thereafter, ten individual HpTFs were measured per subject. To *a priori* account for potential positional variance in the transfer functions, subjects were instructed to move their head to the left and right in between individual headphone measurements. After all measurements, which took about 30 minutes, the investigator removed the microphones without changing the position of the headphones.

2.4. Post-Processing

In a first step, times-of-flight were removed from the BRIRs by means of onset detection and ITDs were calculated and stored separately. ITDs were reinserted in real time during the listening test, avoiding comb-filter effects occurring in dynamic auralization with non-time-aligned BRIRs and reducing the overall system latency [10]. Secondly, BRIRs were normalized with respect to their mean magnitude response between 200 Hz and 400 Hz.

Due to diffraction effects BRIRs exhibit an almost constant magnitude response in this frequency range making normalization especially robust against measurement errors and low-frequency noise. In a last step, BRIRs were truncated to 44100 samples with a squared sine fade out.

Individual HpTF compensation filters were designed using a weighted regularized least mean squares approach [18]. Filters of an FFT order 12 were calculated based on the average of ten HpTF per subject. Regularization was used to limit filter gains if perceptually required, the used approach is shortly explained here: HpTFs typically show distinct notches at high frequencies which are most likely caused by anti-resonances of the pinna cavities [19]. 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 [9]. Therefore, individual regularization functions were composed after manually fitting one or two parametric equalizers (PEQs) per ear to the most disturbing notches. The compensated headphones approached a *target band-pass* consisting of a 4th order Butterworth high-pass with a cut-off frequency of 59 Hz and a 2nd order Butterworth low-pass with a cut-off frequency of 16.4 kHz.

Finally, presentations of the real loudspeaker and the binaural simulation had to be matched to evoke equal loudness impressions. If assuming that signals obtained via individual binaural synthesis closely resemble those obtained from loudspeaker reproduction (cf. Fig. 3), loudness matching can be achieved by simply matching the RMS-level of simulation and real sound field. Hence, matching was pursued by adjusting the RMS-level of five second pink noise samples recorded from loudspeakers and headphones while the subject's head was in the frontal reference position. To account for the actual acoustic reproduction paths in the listening test, prior to loudness-matching, the headphone recordings were convolved with the frontal incidence BRIRs and the headphone compensation filter whereas the loudspeaker recordings were convolved with the target band-pass.

2.5. Test Design

The ABX test paradigm as part of the N-AFC test family provides an objective, criterion-free and particularly sensitive test for the detection of small differences [20], and thus seems appropriate also for a test on the authenticity of virtual environments. ABX-testing involves presenting a test stimulus (A), a hidden reference stimulus (B) and an open reference stimulus (X). Subjects may either succeed (correct answer) or fail (incorrect answer) to identify the test stimulus. Being a Bernoulli experiment with a (2AFC) guessing rate of 50%, the binomial distribution allows the calculation of exact probabilities for observed detection rates enabling tests on statistical significance.

If ABX tests are used to prove the authenticity of simulations, one should be aware that this corresponds to proving the null hypothesis H_0 (i.e., proving equality of test conditions). Strictly speaking, this proof cannot be given by inferential statistics. Instead, the approach commonly pursued is to establish empirical evidence that *strongly supports* the H_0, e.g. by rejecting an alternative hypothesis H_1 stating an effect of irrelevant size, e.g. a minimal increase of the empirical detection rate above the guessing rate (i.e., negating a minimum-effect hypothesis [21]).

When testing a difference hypothesis H_1, two kinds of errors can be made in the final decision: The type 1 (alpha) error refers to the probability of wrongly concluding that there was an audible difference although there was none. The type 2 (beta) error is made, if wrongly concluding that there was no audible difference although indeed there was one. The test procedure (i.e. the number of AFC decisions requested) is usually designed to achieve small type 1 error levels (e.g. 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'). In order to design a fair test we first decided about a practically meaningful effect size to be rejected and then aimed at balancing both error levels in order to statistically substantiate both the rejection and the acceptance of the null hypothesis, i.e. the conclusion of authenticity.

For the current listening test, a number of 24 trials was chosen per subject and for each test condition (i.e., one combination of source direction and stimulus type), ensuring that for 18 or more correct answers, the H_0 ($p_{\text{corr.}} = 0.5$) can be rejected, while for less than 18 correct answers, a specific H_1 of $p_{\text{corr.}} = 0.9$ can be rejected for one test condition, both at equal (i.e., fair) type 1 and type 2 error levels. The chosen statistical design also accounted for the fact that each subject had to conduct 4 repeated tests (i.e. error levels of 5% for individual tests were established by suitable Bonferroni correction). The rather high detection rate of $p_{\text{corr.}} = 0.9$ chosen to be rejected corresponds to our expectation that even small differences would lead to high detection rates, considering the very sensitive test design and the trained subjects available.

2.6. Test Procedure

Nine subjects with an average age of 30 years (6 male, 3 female) participated in the listening test, 3 of them were fairly and 6 of them highly experienced with dynamic binaural synthesis. No hearing anomalies were reported and all subjects had musical background (average 13 years of education). They could thus be regarded as expert listeners.

During the listening test three buttons (A/B/X) were displayed on the screen. Audio playback started, if the one of the buttons on the MIDI interface was pressed. To give the answer "A equals X", the corresponding button had to be pressed and held for a short time. Subjects could take their time at will and repeatedly listen to A, B and X before answering, controlling all interaction with the tactile MIDI interface.

Two audio contents were used: a pulsed pink noise (0.75 s noise, 1 s silence, 20 ms ramps) and an anechoic male speech recording (5 s). The latter was chosen as a familiar 'real-life' stimulus, while noise pulses were believed to best reveal potential flaws in the simulation. Further, 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. As mentioned already, four ABX tests were conducted per subject (2 sources x 2 contents) each consisting of 24 trials. The presentation order of content and source was randomized and balanced across subjects. On average, the test took about 45 minutes. To avoid a drift in head position, subjects were instructed to move their head back to the reference position once between each trial and to keep the head's orientation at approx. 0° elevation throughout the test.

Dynamic auralization was realized using the fast convolution engine *fWonder* [13] in conjunction with an algorithm for real-time reinsertion of the ITD [10]. *fWonder* was also used for

applying (a) the HpTF compensation filter and (b) the loudspeaker target band-pass. The playback level for the listening test was set to 60 dB(A). BRIRs used in the convolution process were dynamically exchanged according to the subjects' current head-above-torso orientation, and playback was automatically muted if the subject's head orientation exceeded 35° azimuth.

2.7. Physical Verification

Prior to the listening test, acoustic differences between test conditions were estimated based on measurements with the FABIAN dummy head. Therefore, FABIAN was placed on the chair and BRIRs and HPTFs were measured and post-processed as described above. In a second step, BRIRs were measured as being reproduced by the headphones and the simulation engine described above. Differences between simulation and real sound field for the left ear and source 1 are depicted in Fig. 3.

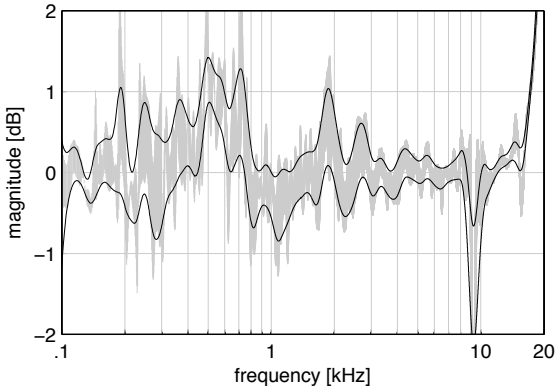

Figure 3: *Differences between binaural simulation and real sound field for source 1 and left ear. The grey area encloses the range of differences observed for all head-above-torso orientations between ±34°. For ease of interpretation, the range of differences is shown again after applying 6th octave smoothing (black lines).*

At a notch frequency in the HpTF at 10 kHz, differences reached up to 6 dB. However, this was assumed to be perceptually irrelevant since the bandwidth of the notch was less than a 10th octave. Above 3 kHz differences were in a range of ±0.5 dB. Somewhat larger and presumably audible deviations of up to ±2 dB were observed between 100 Hz and 3 kHz which were potentially caused by time variance of electro-acoustic transducers. Altogether, Fig. 3 shows comparable error patterns as Fig. 7b in Moore et al. [6].

3. RESULTS

Results of the ABX listening test are summarized in Fig. 4 for all subjects. A clear difference in detection performance was found between contents: While for the pulsed noise subjects were able to discriminate simulation and real sound field (all individual tests were statistically significant, see sect 2.5. for the description of the statistical test), for the speech stimulus about half of them were not (55% significant tests). This increased uncertainty is also reflected in larger variance across subjects. Moreover, a tendency for higher detection rates ($p_{\text{corr.}}$) was found for source 2 ('s2') compared to source 1 ('s1'). Although statistical analysis of detectability was conducted on the level of individual subjects, observed average detection rates are given for better comparability to earlier studies: $p_{\text{corr. noise s1}} = 0.978$, $p_{\text{corr. noise s2}} = 0.991$, $p_{\text{corr. speech s1}} = 0.755$, and $p_{\text{corr. speech s2}} = 0.829$.

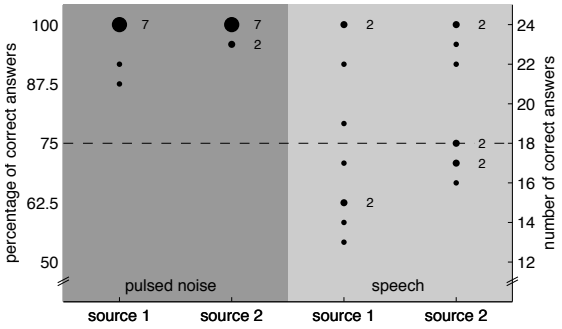

Figure 4: *Listening test results of nine subjects and for each test condition. Dots indicate percentage/number of correct answers for each tested condition; singular numbers indicate subjects with identical detection results. Dots on or above the dashed line indicate statistically significant differences.*

Differences between stimuli could also be found when comparing the average duration needed for making decisions (significantly higher for speech: 38 s vs. 15 s, $p < 0.01$, Wilcoxon signed rank test for dependent samples). Furthermore, increased head movements were found for speech (interquartile range 20° vs. 8° azim., $p < 0.01$, Wilcoxon signed rank test for dependent samples), indicating an extended search behavior adopted by subjects.

During auralization, BRIRs were selected solely based on the subjects' head-above-torso orientation. Hence, unobserved differences in the remaining degrees of freedom (x, y, z, elevation, lateral flexion) might have caused audible artifacts. Therefore, head tracker data were recorded and used for a post hoc analysis of deviations between head position during binaural measurements and ABX tests: For x, y, z coordinates, deviations were found to have been smaller than 1 cm for 95% of the time and never exceed 2 cm which is well within limits given by Hiekkanen et al. [17]. Differences in head elevation (tilt) and in lateral flection (roll) rarely exceeded 10° and were below 5° for 90% of the time. This may have caused audible artifacts occasionally [1], but a systematic influence on the results is unlikely.

When asked for the qualities of perceived differences between simulation and reality after the listening test, subjects named coloration (7x), slight differences in loudness (2x), and spaciousness (1x). Furthermore, two subjects reported a hissing or resonating sound in the decay of the noise pulses.

4. DISSCUSSION AND OUTLOOK

In the present study we assessed whether a state-of-the-art individual dynamic binaural simulation of an echoic environment can still be discriminated from the corresponding real sound field (test of 'perceptual authenticity'). To this end, measurement and post-processing of individual binaural transfer functions was demonstrated to be feasible within a reasonable amount of time, while obtaining a sufficient SNR and avoiding excessive test subject fatigue. Further, listening tests were conducted immedi-

ately after the measurements (i.e., – due to the minimization of deviations caused by time variability – resembling a best case scenario when aiming at proving authenticity) using a sensitive ABX test paradigm.

In accordance with earlier studies, we found that for a pulsed pink noise sample all subjects could reliably detect a difference between reality and simulation (individual detection rates between 87.5% and 100%). In case of the speech sample, however, only about half of the subjects still perceived a difference (individual detection rates between 54% and 100%). The higher detectability for the noise stimulus can be explained by its broadband and steady nature, supporting the detection of coloration, which, according to the subjects, was perceived as the major difference. Further, in considering this, also the mentioned loudness differences might be related to remaining spectral deviations.

Furthermore, higher detection rates were observed for source 2 as compared to source 1. These could be explained by occasionally observed slight discontinuities in the extracted ITD, most probably due to lower SNR at the contralateral ear. Additionally, low SNR might have led to larger measurement errors potentially perceivable as coloration.

Further, a tendency for interaction between source and type of stimulus was observed, as across all subjects, detection rate was by far lowest for source 1 and the speech stimulus ($p_{corr.s1.noise}$ = 75.5%). The observed value indicates that for this condition the group's detection performance was at threshold level (discrimination between simulation and reality in 50% of the cases, equalling 75% in a 2AFC paradigm).

On overall, the observed detection rates were higher than those reported in previous studies, although the precision of the binaural reproduction was comparable [6]. Hereby, our test design allowing subjects to switch at will between stimuli before making final decisions, may be assumed to be much more sensitive to small flaws of the simulation than sequence-based presentations applied in previous studies. This is also indicated by the fact that six subjects reported to have felt to be merely guessing although four of them produced significant detection results for one source of the speech stimulus. In addition, results indicate that it is still more demanding to realize an authentic interactive real time simulation as compared to static auralization. This was somehow expectable as extended abilities of a simulation naturally go together with extended potential for perceptual issues (e.g., with respect to crossfading, latency, or spatial discretization).

Moreover, and in contrast to former studies, our test included simulating a reverberant environment. Future tests which are planned to be conducted in an anechoic chamber and a concert hall will reveal whether the simulation of reverberant environments resembles a specific challenge.

The 'hissing' sound perceived by two subjects might be an artefact related to slightly mistuned headphone filters, indicating the potential for future improvements of our simulation as e.g. with respect to perceptually more robust headphone filter design. Further, an optimization of individual ITD modelling appears advisable and will be pursued in the future.

5. SUMMARY

A test of authenticity was conducted for the first time for a dynamic individual binaural simulation. Results showed that when by applying a sensitive test design the simulation was always clearly distinguishable from the real sound field, at least for critical sound source positions and if presenting noise bursts. However, for male speech, resembling a typical 'real-life' audio content and for a non-critical source position, half the subjects failed to reliably discriminate between simulation and reality, and averaged across subjects performed at threshold level.

6. ACKNOWLEDGMENTS

This work was funded by the German Research Foundation (DFG WE 4057/3-1).

7. REFERENCES

[1] J. Blauert, *Spatial Hearing. The psychophysics of human sound localization*, MIT Press, Revised Edition, Massachusetts, USA, 1997.

[2] R. S. Pellegrini, "A virtual reference listening room as an application of auditory virtual environments," Ph.D. thesis, University Bochum, 2001.

[3] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments," *Acta Acust. united Ac.*, vol. 98, no. 5, pp. 804-810, 2012.

[4] H. Wierstorf, A. Raake, M. Geier and S. Spors, "Perception of focused sources in wave field synthesis," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 5-16, 2013.

[5] E. H. A. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 528-537, 2000.

[6] A. H. Moore, A. I. Tew and R. Nicol, "An initial validation of individualized crosstalk cancellation filters for binaural perceptual experiments," *J. Audio Eng. Soc. (Engineering Report),* vol. 58, no. 1/2, pp. 36-45, 2010.

[7] B. Masiero, "Individualized Binaural Technology. Measurement, Equalization and Perceptual Evaluation," Ph.D. thesis, RWTH Aachen, 2012.

[8] V. Erbes, F. Schulz, A. Lindau and Stefan Weinzierl, "An extraaural headphone system for optimized binaural reproduction," *Fortschritte d. Akustik: Tagungsband d. 38. DAGA [German annual acoustic conference]*, pp. 313-314, Darmstadt, Germany, March, 2012.

[9] A. Lindau and F. Brinkmann, "Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings," *J. Audio Eng. Soc.*, vol. 60, no. 1/2, pp. 54-62, 2012.

[10] A. Lindau, J. Estrella and S. Weinzierl, "Individualization of dynamic binaural synthesis by real time manipulation of the ITD," in *Proc. 128th AES Convention*, London, UK, May 22-25, 2010.

[11] A. W. Mills, "On the minimum audible angle," J. Acoust. Soc. Am., vol. 30, no. 4, pp. 237-246, 1958.

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

[13] A. Lindau, T. Hohn and S. Weinzierl, "Binaural resynthesis for comparative studies of acoustical environments," in *Proc. 122th AES Convention, Convention Paper 7032*, Vienna, Austria, May, 2007.

[14] W. M. Hartmann and A. Wittenberg, “On the externalization of sound images,” *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3678-3688, 1996.

[15] A. Lindau and S. Weinzierl, “On the spatial resolution of virtual acoustic environments for head movements on horizontal, vertical and lateral direction,” in *Proc. EAA Symposium on Auralization*, Espoo, Finland, June 15-17, 2009.

[16] S. Müller and P. Massarani, “Transfer function measurement with Sweeps. Directors's cut including previously unreleased material and some corrections,” *J. Audio Eng. Soc. (Original Release)*, vol. 49, no. 6, pp. 443-471, 2001.

[17] T. Hiekkanen, A. Mäkivirta and M. Karjalainen, “Virtualized listeneing tests for loudspeakers,” *J. Audio Eng. Soc*., vol. 57, no. 4, pp. 237-251, 2009.

[18] S. G. Norcross, M. Bouchard and G. A. Soulodre, “Inverse Filtering design using a minimal phase target function from regularization,” in *Proc. 121th AES Convention, Convention Paper 6929*, San Francisco, USA, October 5-8, 2006.

[19] 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.*, vol. 132, no. 6, pp. 3832-3841, 2012.

[20] L. Leventhal, “Type I and type 2 errors in the statistical analysis of listening tests,” *J. Audio Eng. Soc.*, vol. 34, no. 6, pp. 437-453, 1986.

[21] 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.*, vol. 84, no. 2, pp. 234-248, 1999.