OPEN ACCESS Freely available online

D. Ackermann, F. Fiedler, F. Brinkmann, M. Schneider, and S. Weinzierl, “On the Acoustic Qualities of Dynamic Pseudobinaural Recordings”
J. Audio Eng. Soc., vol. 68, no. 6, pp. 418–427, (2020 June.).
DOI: https://doi.org/10.17743/jaes.2020.0036

On the Acoustic Qualities of Dynamic Pseudobinaural Recordings

DAVID ACKERMANN[1], FELICITAS FIEDLER[1], FABIAN BRINKMANN[1], *AES Associate Member*,
(david.ackermann@tu-berlin.de)
MARTIN SCHNEIDER[2], *AES Member*, AND STEFAN WEINZIERL[1]

[1] *Technische Universität Berlin, Berlin, Germany*
[2] *Georg Neumann GmbH, Berlin, Germany*

The motion-tracked binaural (MTB) technique allows the dynamic, pseudobinaural rendering of spatial sound scenes recorded by a circular array of microphones on a rigid sphere. The system provides a multichannel live audio transmission from which a head-related signal with approximated interaural time and level differences can be derived and played via headphones, head tracking, and a corresponding rendering software. The latter is mainly calculating imperceptible interpolation between channel pairs during head movements. This contribution evaluates the potential of this format for the creation of virtual acoustic environments. Based on the technical realization of a 16-channel MTB array with omnidirectional diffuse field-corrected electret condenser microphone capsules, the *plausibility* of 8 and 16-channel recordings was tested against a physical sound source. Furthermore, the sound quality of the pseudobinaural rendering was assessed based on different items of the Spatial Audio Quality Inventory (SAQI) compared to a true dynamic binaural reference. The results show that the overall plausibility of the MTB signal with optimal interpolation is close to the reference. Even if there are small differences with respect to tone color and spatial sound source attributes, the degree of externalization and even the perceived source elevation were, despite the absence of pinna cues, well comparable to the true binaural reference.

0 INTRODUCTION

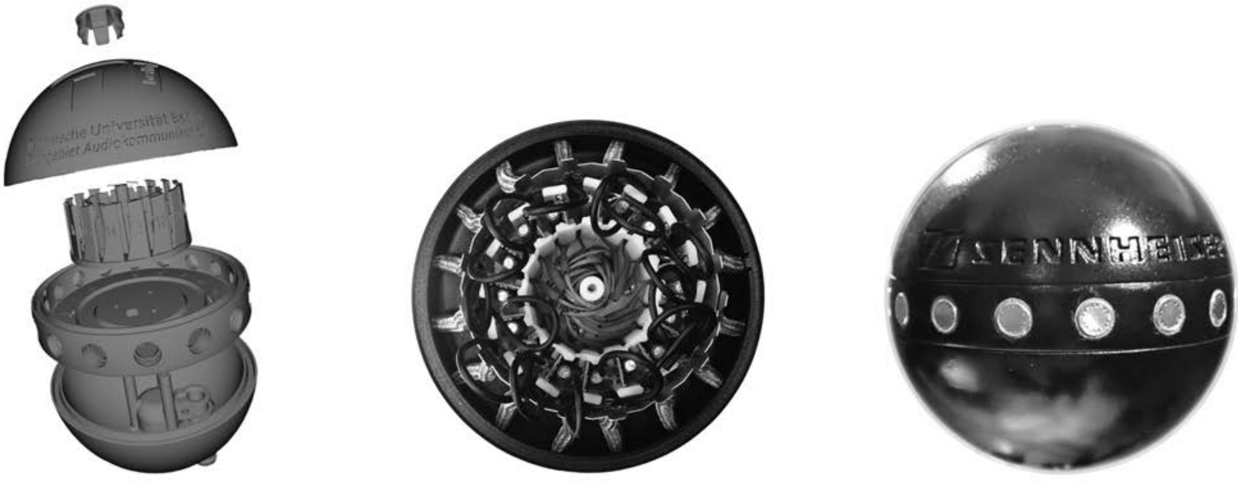
The motion-tracked binaural (MTB) technique is a method for the pseudobinaural recording and rendering of spatial sound scenes [1]. The recording device consists of a rigid sphere with a diameter based on the size of a human head and an equidistant array of microphones on the equator of the sphere. During playback, dynamic head-related signals can be obtained by interpolating between pairs of opposing microphones according to the head orientation of the listener. The MTB method thus enables a dynamic encoding and transmission of acoustic environments in real time, not only accounting for head movements of the listener but also for movements of sound sources with six degrees of freedom. This possibility of an efficient real time streaming of audio scenes makes pseudobinaural recordings an alternative not only to classical dynamic binaural synthesis, which requires one binaural room impulse response (BRIR) for each head orientation, source position, and source orientation, but also to a binaural encoding using a spherical microphone array, which requires a significantly higher number of microphones [2, p. 225f.]. MTB signals, however, fail to account for the influence of the pinna, with possible consequences for timbral properties, the localization of elevation, and high-frequency interaural level differences (ILDs).

The aim of this study is thus to evaluate the perceptual quality of pseudobinaural recordings, realized with a 16-channel MTB microphone developed in-house, against a true binaural reference. To this end, we tested the *plausibility* of MTB recordings as an integral quality measure [3] and conducted a qualitative analysis of artifacts based on the Spatial Audio Quality Inventory (SAQI) [4].

1 METHOD

1.1 Array Design and Signal Processing

The MTB microphone used is a spherical rigid microphone array whose radius of $d = 8.75$ cm was determined by a least mean square fit between measured interaural time differences (ITDs) and spherical head ITDs calculated



(a) 3D model (b) top view (interior) (c) side view (assembled)

Fig. 1. The custom-made 3D model (a) of the motion-tracked binaural (MTB) for Selective Laser Sintering (SLS), the interior from top view (b)m and the assembled microphone array from side view (c).

with the Woodworth formula [5]. The array was 3D printed from a thermoplastic material using Selective Laser Sintering (SLS) based on a custom made 3D model. Sixteen hand-matched omnidirectional Sennheiser KE14 electret condenser capsules (free field sensitivity of 32 mV/Pa, equivalent sound pressure level of 15 dB(A), and a maximum sound pressure level of 130 dB(A) for a total harmonic distortion (THD) of 1% at 1 kHz) were mounted flush with the surface on the horizontal circumference of the sphere. The capsules have a largely flat frequency response with a tolerance of ±3 dB up to 16 kHz and maximum differences between capsules smaller than 2 dB.

The integrated voltage converter allows the operation of the microphone array with a phantom power of 12 to 48 V provided by two 8-channel Presonus Digimax DP88 preamplifiers that are also used for A/D conversion and were connected to an RME Fireface via ADAT. Fig. 1 shows the custom-made 3D model and the interior view of the MTB, as well as the assembled microphone array.

Due to the discrete spatial sampling of the sound field, it is necessary to interpolate between the signals of two adjacent capsule pairs with weights determined by the current head position in order to achieve a spatially continuous reproduction of the two ear signals during head rotation. Algazi et al. [1] suggested five different interpolation methods, with the *Two-band Spectral-Interpolation Restoration* algorithm performing best when evaluated with different numbers of microphone capsules (8, 16, 24, 32) and audio contents (noise, music, speech) [6]. With this method, the low-frequency component is linearly interpolated directly in the time domain, and high frequencies are interpolated in the frequency domain using short-time Fourier transforms (128 samples, 75% overlap, Hanning windowed). While the high-frequency magnitude response is obtained by linear interpolation, the phase response is taken from the nearest neighbor to avoid comb-filter coloration [1]. The crossover frequency between time domain and frequency domain interpolation depends on the number of microphone capsules N and is given by

$$f_{\text{x-over}} = \frac{Nc}{8\pi r}, \tag{1}$$

where $c \approx 343$ m/s is the speed of sound and $r = 8.75$ cm the MTB radius [1]. We thus obtain crossover frequencies of ca. 1.25 kHz ($N = 8$) and 2.5 kHz ($N = 16$). Real-time interpolation and playback of the MTB signals is done by the MTB renderer [6], which is a Linux-based *Jack Audio Connection Kit* (JACK) client controlled by *Open Sound Control* (OSC) messages and a Polhemus Patriot head tracker.

To achieve an uncolored timbre in reverberant environments, a diffuse field equalization filter was designed, considering the fact that MTB recordings will be played back via headphones that are most likely to already exhibit a diffuse field like target function (cf. [7] and headphone transfer functions in [8]). To this end, the frequency response of one MTB capsule was measured in the anechoic chamber at Georg Neumann GmbH in Berlin for 64 positions on the horizontal plain between 0° (front) and 180° (back). The area weighted root mean square averaged diffuse field transfer function was then calculated over a virtually completed full spherical sampling grid by exploiting the point-symmetric design of the MTB and relying on the matching of the capsules. The raw equalization filter was obtained by calculating the inverse, and the final filter was compressed with a ratio of 4:1 and a threshold of 5 dB to avoid excessive boost at high frequencies that would considerably amplify the microphone noise (cf. Fig. 2). This gain restriction was found to be a good compromise between compression induced coloration and gain induced noise in an informal listening test conducted by the authors.

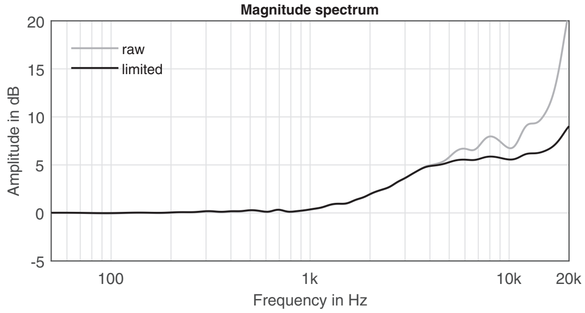


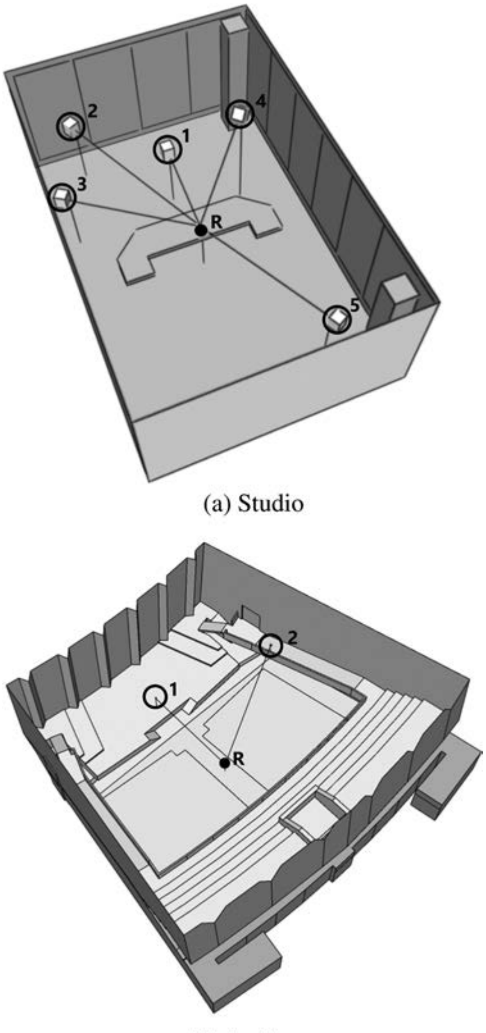
Fig. 2. Raw area weighted and route mean square averaged diffuse field equalization filter (gray) and final filter compressed with a 4:1 ratio and a threshold of 5 dB (black).

1.2 Assessment of Plausibility

The *plausibility* of a virtual acoustic environment can be defined as a simulation "in agreement with the listener's expectation towards a corresponding real event" [3, p. 804]. It was assessed by playing short audio examples of either real or simulated sound sources to the subjects, who had to decide whether they assumed the sound came from a real loudspeaker or from a binaural simulation of the loudspeaker without being able to directly compare the two events. The listeners were only told that real and simulated stimuli were approximately evenly distributed across 100 trials. From the answers, the sensitivity d' and the bias β can be calculated based on Signal Detection Theory analysis [3]. The sensitivity describes the subjects' ability to distinguish simulation and reality with respect to an inner reference, and the bias indicates the tendency towards answering with *yes* or *no*. This conceptual separation of sensitivity and bias makes the plausibility test a double-blind and criterion-free procedure. The sensitivity values can be converted to detection rates P_c of a corresponding two-alternative forced choice (2AFC) test, with $P_c = \Phi(d'/\sqrt(2))$ and $\Phi(z)$ the cumulative standard normal distribution, yielding $p_{2AFC} = 0.5$ (guessing) for $d' = 0$ and $p_{2AFC} > 0.5$ for $d' > 0$. After calculating d' for every subject, a *t* test can be conducted to determine if the group sensitivity is significantly lower than a predefined critical value d'_{min} (minimum effect test [9] with the null hypothesis $H_0 : d' = d'_{\text{min}}$). Lindau and Weinzierl [3] chose a critical value of $d'_{\text{min}} = 0.1777$ corresponding to a detection rate of $p_{2AFC} = 0.55$ that exceeds the guessing rate by only 5% to determine the plausibility of a nonindividual, dynamic binaural synthesis. Based on this minimum effect hypothesis, an optimal sample size of $N_{\text{opt}} = 1071$ trials was set by tolerating a type I error of 25% and a type II error of 5% resulting in 11 subjects and 1,100 trials in total (100 per subject).

The listening test took place in the Electronic Studio of the Audio Communication Group, with a reverberation time of $T_m = 0.28$ s and a volume of $V = 140$ m^3. Five Genelec 8020c studio monitors were positioned around a central listening spot according to Fig. 3 and Table 1.

Prior to the listening test, 16-channel MTB room impulse responses (RIRs) were measured for the five loudspeakers at the listening position. To generate the simulated stimuli for the MTB renderer [6], each 16-channel RIR was convolved with 20 different audio contents ranging from male and female speech in different languages to recordings of solo instruments and extracts of pop songs with a length of 3–6 s. An extraaural headphone (BK2/11) [10] was used to present the MTB signals that allowed the presentation of the real signals with negligible influence on



(a) Studio

(b) Audimax

Fig. 3. Loudspeaker positions (circles) and listening position R (black dot) in the Electronic Studio (a) and Audimax (b). Speakers 1 and 4 of the Studio were used in both listening tests.

Table 1. Loudspeaker positions in the Electronic Studio and Audimax. Speakers 1 and 4 of the Studio were used in both listening tests.

	Speaker	Azimuth	Elevation	Distance
Studio	1	0°	0°	1.50 m
	2	28°	6°	3.69 m
	3	60°	8°	2.65 m
	4	−45°	48°	1.65 m
	5	−126°	7°	3.16 m
Audim.	1	0°	0°	12.7 m
	2	−70°	17°	16.6 m

the external sound field [11, Fig. 2]. For tracking the head movements, a Polhemus Patriot head tracking system was used. The headphones were equalized towards the diffuse field transfer function of the FABIAN head and torso simulator provided in Brinkmann et al. [8], and the MTB signals were filtered with the inverse diffuse field filter (cf. Fig. 2). An additional 4th-order Butterworth high pass was applied at 40 Hz for low frequency noise rejection. The playback level of real and simulated stimuli was adjusted to an equal loudness by the authors.

At the beginning, the participants were informed about the nature of the listening test and familiarized with the user interface. The subjects were encouraged to move their heads in the horizontal plane without restriction during the listening test in order to exploit the full potential of the dynamic rendering. A chair with a neck rest allowing head rotations to the left and right was used to avoid positional shifts during the experiment. After listening to a stimulus exactly one time, the subjects answered the question "Was the audio played back by a real loudspeaker (yes/no)" and proceeded to the next trial. A custom Matlab user interface was used to randomize the stimulus presentation, acquire the subjects' answers, and to control the MTB renderer via OSC messages. In order to prevent memory effects that could influence the response bias of the subjects, each of the 20 audio contents was played back from each of the five source positions exactly once during the 100 test trials.

1.3 Qualitative Evaluation

A qualitative evaluation was conducted based on the Spatial Audio Quality Inventory [4] that contains 48 items, six of which were selected based on their relevance assessed in a pretest for the MTB: *tone color* (darker–brighter), *vertical direction* (shifted down–shifted up), *distance* (closer–more distant), *externalization* (more internalized–more externalized), *localizability* (more difficult–easier), and *naturalness* (lower–higher). The quality of 8 and 16-channel MTB renderings was assessed in a four-factorial fully repeated measures test with the factors *channels* (8, 16), *room* (dry, wet), *source position* (frontal, top right), and *content* (male speech, white noise), resulting in eight test conditions per SAQI item. The Electronic Studio ($T_m = 0.28$ s, $V = 140$ m^3) and the largest lecture hall of the TU Berlin (Audimax ($T_m = 2.1$ s, $V = 8500$ m^3) were chosen as *dry* and *wet* environments, and two sources were selected to assess the rendering for a typical case (frontal) and a critical case (top right), considering that pinna dependent elevation cues are not present in the MTB signals (cf. Fig. 3 and Table 1).

To provide a true binaural reference for the MTB, BRIRs of the FABIAN head and torso simulator [12] were measured at the positions shown in Fig. 3 and Table 1 for head orientation to the left and right between ±40° with a resolution of 1°. In order to achieve the best possible auralization of the non-individual BRIRs, the interaural time difference in the binaural signals was individualized separately for each participant by ITD extraction and manipulation based on the measured intertragus distance [13]. A 40 Hz, 4th-order Butterworth high-pass was applied to all signals to reject low-frequency noise. To avoid unintended coloration between the MTB and BRIR renderings, a diffuse field equalization was done by means of convolution with the inverse filters shown in Fig. 2 and provided by Brinkmann et al. [8, file: `FABIAN_CTF_measured_inverted_smoothed.sofa]`. The transfer function of the Sennheiser HD800S headphones that were used for audio playback was not compensated to provide a consumer-like listening situation. The combination of diffuse field-compensated renderings with approximately diffuse field-equalized headphones (cf. transfer functions in [8]) proved to provide a natural tone color in informal listening tests conducted by the authors. A Polhemus Patriot head tracker was used for tracking the subjects' head orientations.

The experiment took place in the Electronic Studio at TU Berlin. The subjects were first informed about the nature of the experiment, introduced to the definition of the SAQI items (cf. [4, Table 1]), and familiarized with the test procedure in a short training session. The subjects' answers were acquired through a Matlab-based user interface (cf. [14]) that showed the name of the current current quality and a rating slider with the scale labels (e.g., *tone color:* darker–brighter). Five lines at equally spaced distances were provided next to the sliders for orientation. For the *vertical direction*, a text field for entering the perceived difference in degrees was provided instead of the quasicontinuous sliders. Two buttons labeled *A* and *B* were shown below the slider to start the playback of the reference (*A*) and MTB signals (*B*). The subjects were not informed about the assignment of the test conditions to the buttons and were instructed to rate the quality of *B* with respect to *A*, to listen to the stimuli as often as and in any order they wanted, and to move their heads within the range of ±40° to the left and right. The order of items and test conditions was randomized across subjects. The user interface also sent OSC messages to control the playback of the MTB renderer and a customized version of the SoundScape Renderer that loaded a Spatially Oriented Format for Acoustics (SOFA) file [15] containing the reference BRIRs. The listening test took approximately 45 minutes including instructions and training.

2 RESULTS

2.1 Plausibility

Separate groups of 11 subjects each were used to evaluate the plausibility of 8 and 16-channel MTB rendering. The between-subjects design was chosen to avoid familiarization with the stimuli that might have increased the detection rate if one subject took part in two plausibility tests. The participants had an average age of 29 years (16 male, 6 female) and no self-reported hearing impairments. 16 subjects had several years of musical education and 8 participants were experienced with listening tests using dynamic binaural synthesis.

Individual and group-averaged sensitivities d' and biases β are given in Fig. 4 and Fig. 5, along with 90% bootstrap confidence intervals ($n_{\text{boot}} = 2000$ samples, nonparametric

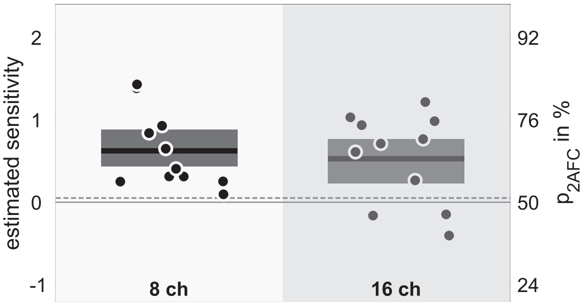


Fig. 4. Results of the test for plausibility. Estimated individual sensitivities d' (left y-axis) and corresponding two-alternative forced choice (2AFC) detection rates p_{2AFC} (right y-axis) are given by the points (offset in horizontal direction to improve readability). The gray dashed line shows the mean value of $p_{2AFC} = 0.514$ ($d'_{mean} = 0.0512$) achieved by Lindau and Weinzierl [3] for a true data-based dynamic binaural synthesis. Correct answers of 50% denote guessing. The boxes show the group mean and 90% bootstrapped confidence intervals (CIs).

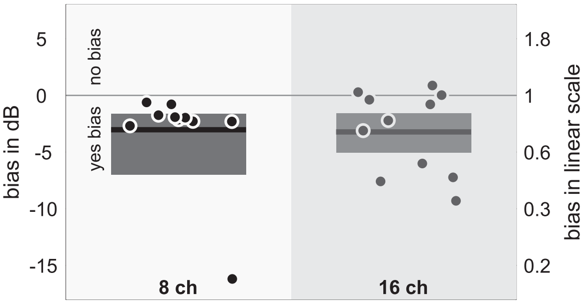


Fig. 5. Results of the test for plausibility. Estimated individual biases β in dB (left y-axis) and in linear scale (right y-axis) are given by the points (offset in horizontal direction to improve readability). The boxes show the group mean and 90% bootstrapped confidence intervals (CIs).

resampling, bias-corrected and accelerated CI calculation [16]). An average sensitivity of $d'_{\text{mean}} = 0.62$ ($p_{2AFC} = 0.66$) was observed for an 8-channel MTB rendering and $d'_{\text{mean}} = 0.53$ ($p_{2AFC} = 0.64$) for a 16-channel rendering. A two-sided *t* test showed that there is no significant difference between 8 and 16-channel MTB playback ($p = 0.66$, type I error 0.05). Prior to the *t* test, a Kolmogorov–Smirnov test confirmed the underlying assumption of normality for 8 ($D(11) = 0.23$, $p = 0.12$) and 16-channel ($D(11) = 0.20$, $p = 0.2$) MTB playback. The β values shown in Fig. 5 suggest a slight but statistically significant bias of the subjects towards believing in the realness of the stimuli.

2.2 Qualitative Evaluation

A total of 30 subjects (21 male, 8 female, 1 nonbinary, 29 years of age on average) participated in the experiment, of which 23 had several years of musical education and 17 had participated in listening tests with dynamic binaural synthesis before.

The results of the SAQI test are shown in Fig. 6 for all test conditions. They describe the perceived difference between the MTB and the true binaural reference whereby a value 0 indicates no difference and a value of ±1 a very large difference. Because the ratings were not normally distributed in all cases (Shapiro–Wilk test), ratings are described by their median and 95% bootstrap confidence intervals ($n_{boot} = 2{,}000$ samples, nonparametric resampling, bias-corrected and accelerated CI calculation [16]).

Speech signal: The differences between the MTB signals and the dynamic binaural synthesis show a small change in tone color for the speech stimulus, with median values of the test signals in the range of [0; 0.24] compared to the reference. Minor differences in source position were perceived in the vertical plane (elevation), with a maximum position difference of 5° (median value, shifted up) for the elevated source in the Audimax. All other median values are at 0. Concerning the distance, the median values are at the same level as the binaural reference for all conditions. Only the top right source in the Studio stands out slightly. It was perceived a little closer both with 16-channel (median value −0.19) and with 8-channel MTB playback (median value −0.2). The same tendency can be observed in the degree of externalization, where the elevated source in the Studio was perceived as more internalized both with 16-channel (median value −0.1) and with 8-channel MTB playback (median value −0.19). The localizability of the MTB test signals was perceived as equal or even more precise as the reference for each condition. The frontal source in the Audimax was even rated with a median value of 0.14 as easier to locate. In terms of the naturalness, the participants could perceive no difference to the binaural synthesis on average. The median values are 0 for all cases.

Noise signal: Larger differences were perceived in the noise signal. In tone color, all MTB signals were perceived as brighter than the reference, with median values between 0.4 and 0.43. Only minor effects were noticed in the elevation of the sources, with a median position difference of 5° (frontal source in the Audimax, 8- and 16-channel MTB playback) and 1.5° (frontal source in the Studio, 8-channel playback) shifted up. Regarding distance, the source positions in the noise signal were perceived as closer to the listening position than in the speech signal (median values of ca. −0.2), in particular for the elevated source in the Studio, with a median value of −0.33 for both 8 and 16 MTB channels. This source was also perceived as more internalized, with a value of −0.27 for the 16-channel playback and −0.35 for the 8-channel playback. The same tendency was observed for the frontal source in the Audimax for playback with 8 MTB channels (−0.19) and the elevated source for playback with 16 MTB channels (−0.14). The localizability of the sound sources in the test signals varies slightly more than for the speech signal. While the frontal source in the Audimax could be located more easily than the reference with MTB playback, subjects found it harder to locate the top right source in the Studio (maximum deviation at 0.4 for 16 MTB channels). The other values fluctuate around 0. The assessments of naturalness hardly deviate from the reference. Just the two source positions in the Audimax were perceived as less natural, with median values down to −0.2.

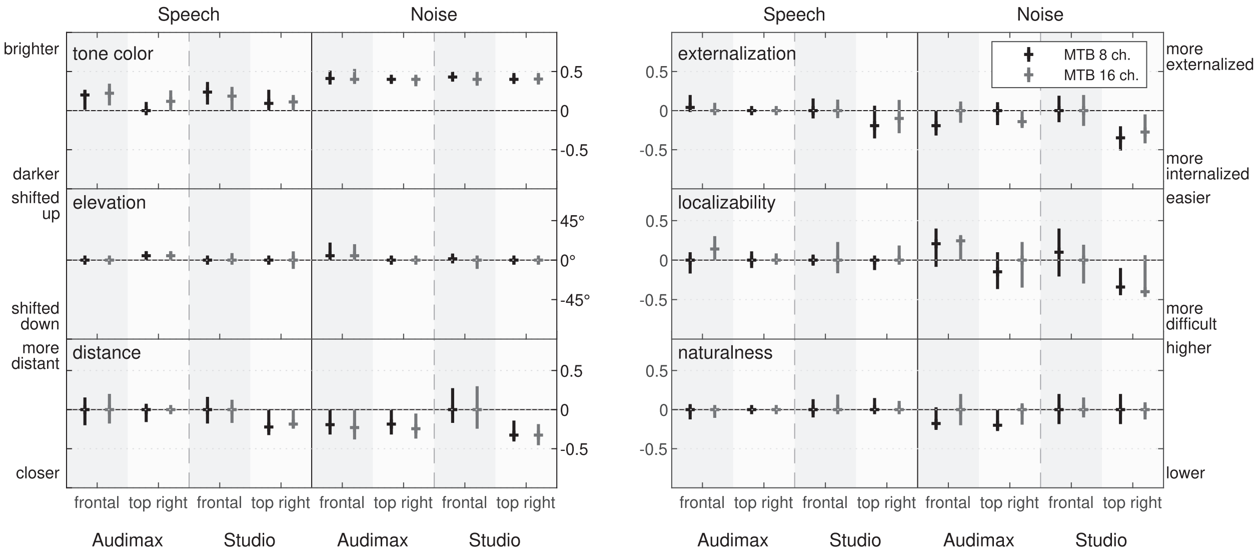


Fig. 6. Differences in specific auditory qualities, measured with attributes of the Spatial Audio Quality Inventory (SAQI), showing the median of differences between the motion-tracked binaural (MTB) and the true binaural reference (horizontal lines) with 95% bootstrap confidence intervals (vertical lines). The ratings were given for speech and noise as audio content and for 8 and 16-channel MTB rendering (from left to right).

For further insights, within-subject differences were analyzed by means of a four-factorial repeated measures multivariate analysis of variance (MANOVA) with the factors *channels*, *content*, *room*, and *position*. Absolute Pearson correlation coefficients between SAQI items were below 0.21, indicating only weak collinearity that agrees with the MANOVA assumptions and a visual inspection of the model residuals showed no obvious deviation from normal distribution. The Mauchly test for sphericity is irrelevant in this case, since the inner subject factors each have only two forms and thus sphericity is given.

The MANOVA showed significant multivariate main effects for three of the four factors considered. This includes *content* (*Pillai's Trace* =0.819, $F(6, 24) = 18.04$, $p < 0.001$, partial $\eta^2 = 0.82$), *room* (*Pillai's Trace* =0.439, $F(6, 24) = 3.13$, $p < 0.05$, partial $\eta^2 = 0.44$), and *position* (*Pillai's Trace* =0.552, $F(6, 24) = 4.93$, $p < 0.01$, partial $\eta^2 = 0.55$), as well as the interactions between *content* × *position* (*Pillai's Trace* =0.398, $F(6, 24) = 2.64$, $p < 0.05$, partial $\eta^2 = 0.4$) and *room* × *position* (*Pillai's Trace* =0.451, $F(6, 24) = 3.28$, $p < 0.05$, partial $\eta^2 = 0.45$). No significant multivariate main effect was observed for the factor *channel* (*Pillai's Trace* =0.132, $F(6, 24) = 0.607$, $p = 0.72$, partial $\eta^2 = 0.132$).

Univariate analyses of variance (ANOVAs) showed that the *content* had a significant effect on tone color ($F(1) = 63.91$, $p < 0.001$, partial $\eta^2 = 0.688$), distance ($F(1) = 17.26$, $p < 0.001$, partial $\eta^2 = 0.373$) and externalization($F(1) = 1.34$, $p < 0.01$, partial $\eta^2 = 0.231$). For the noise signal, post-hoc Bonferroni corrected pairwise comparisons showed that the tone color was perceived significantly brighter (mean 0.38 noise vs. 0.17 speech), the sound sources were significantly closer (-0.18 vs. -0.04), and perceived to be more internalized (-0.17 vs. 0). In addition, *room* only had a significant effect on elevation ($F(1) = 6.30$, $p < 0.05$, partial $\eta^2 = 0.179$). The Bonferroni-corrected pairwise comparisons showed that the sources in the Audimax were perceived to be significantly shifted upwards compared to the Studio (6.7° vs. $-0.5°$). Significant univariate main effects were also observed for the *source position* on color ($F(1) = 10.26$, $p < 0.01$, partial $\eta^2 = 0.261$), elevation ($F(1) = 4.20$, $p = 0.05$, partial $\eta^2 = 0.126$), distance ($F(1) = 8.20$, $p < 0.01$, partial $\eta^2 = 0.261$), externalization ($F(1) = 13.28$, $p < 0.01$, partial $\eta^2 = 0.314$), and localizability ($F(1) = 11.08$, $p < 0.01$ partial $\eta^2 = 0.277$). Post-hoc tests proved that the frontal sources were perceived brighter (0.31 frontal vs. 0.24 top right) and higher than the reference (6° vs. $-0.2°$), whereas the top right source was perceived to be significantly closer to the listener (-0.02 top right vs. -0.05), slightly more internalized (-0.1 vs. 0), and more difficult to localize (-0.1 vs. 0.1). Significant interactions were observed for *content* × *position* for localizability ($F(1) = 6.06$, $p < 0.05$ partial $\eta^2 = 0.173$) and for *room* × *position* for elevation ($F(1) = 5.86$, $p < 0.05$ partial $\eta^2 = 0.168$), distance ($F(1) = 5.80$, $p < 0.05$ partial $\eta^2 = 0.167$) and externalization ($F(1) = 6.06$, $p < 0.05$ partial $\eta^2 = 0.131$). However, all interactions were ordinal and do thus have no influence on the main effects described above.

3 DISCUSSION

The motion-tracked binaural (MTB) recording of a frontal and a lateral sound source in two different acoustical environments with 20 different audio stimuli has been shown to provide a surprisingly high plausibility of the corresponding sound events. If confronted with a real sound source—a physical loudspeaker in the present test—these sound events can be identified as "simulated" with a mean sensitivity of only $d' = 0.53$, corresponding to a 2AFC

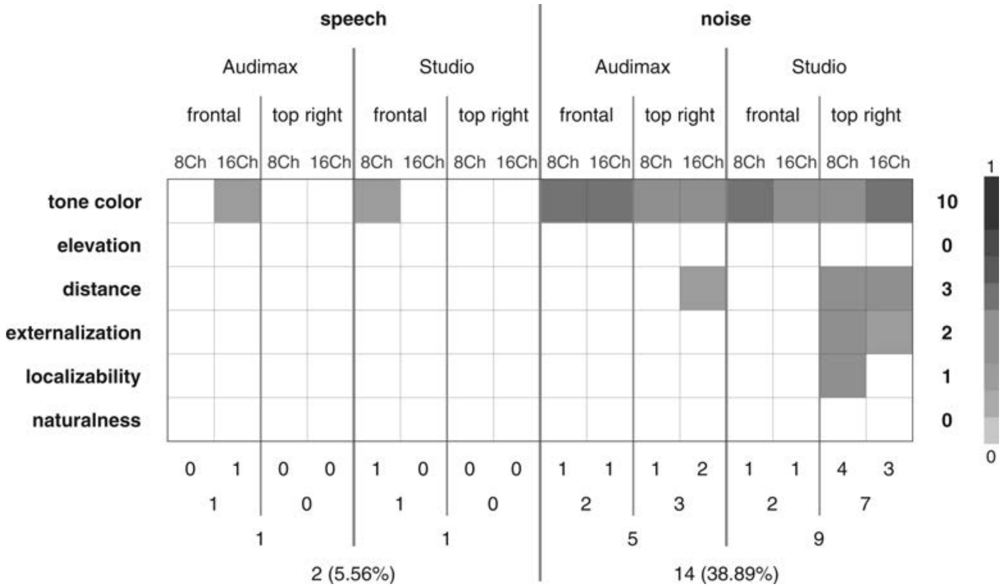


Fig. 7. Results of the Spatial Audio Quality Inventory (SAQI) test. Degree of deviations by audio content, room, source position, number of channels used, and perceptual quality. White areas denote confidence intervals (CIs) overlapping with 0, shaded areas denote CIs not overlapping with 0, in which case the shading denotes the absolute median ratings in the range between 0 and 1 as indicated by the color bar. Numbers indicate the sum of significant deviations across rows and columns.

detection rate of 64%, compared to a guessing rate of 50% for 16-channel MTB rendering. Even if this is less sensitive than the value of $d' = 0.0512$ corresponding to a detection rate of $p_{2AFC} = 0.514$ that was determined for a nonindividual, dynamic binaural synthesis [3], it is remarkable that, despite the lack of pinna cues, the test subjects did not perceive the physical loudspeaker as more plausible than its MTB rendering over headphones in most trials. That the plausibility was unexpectedly high for most of the participants is reflected by the negative bias indicated by the SDT analysis, which results from the fact that they more often considered an MTB rendering to be "real" than the real loudspeaker to be "simulated."

The qualitative evaluation shows both the advantages and the deficits of the MTB encoding. For 16 out of 96 conditions, significant differences between the MTB rendering and the true binaural reference were observed (Fig. 7). Most of these differences were related to tone color and to spatial attributes such as distance, externalization, and localizability.

With regard to tone color, the fact that the MTB was perceived as slightly brighter than the binaural reference can be explained by comparing the diffuse field equalized frequency responses of the FABIAN BRIR and the MTB, as indicated for the frontal and top right sound source in the Studio environment (sources 1 and 4 in Table 1, Fig. 8). For the frontal source position, the frequency response of the MTB is distinctly above the reference between 5 and 10 kHz. For the elevated lateral source, the binaural reference shows larger differences between the ipsilateral (right) and the contralateral (left) than the right and left channel of the MTB between 3 and 9 kHz. Between 10 and 18 kHz, however, the MTB is up to 10 dB above the BRIR for both ears. This deviation of the frequency responses at higher frequencies is due to missing pinna cues, which also results in a lower ILD (Fig. 9). In this context, the diffuse field equalization applied has obviously led to an overemphasis of higher frequencies. A weighted equalization that slightly favors frontal sound incidence might help to reduce these effects for concert-like situations where a frontal sound source does not create a *perfectly* diffuse sound field.

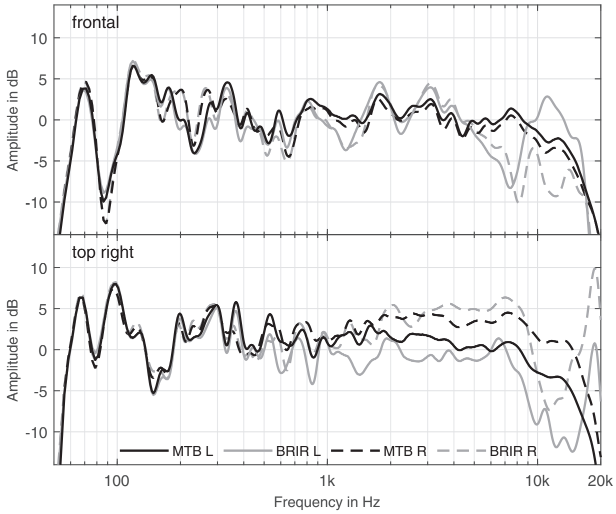


Fig. 8. Diffuse field-compensated frequency responses measured in the studio for the binaural room impulse response (BRIR, gray) and motion-tracked binaural (MTB, black), for the left (solid) and right (dashed) ear with frontal head orientation, and for frontal (top) and elevated lateral (bottom) source position.

It is remarkable that the MTB rendering, despite the lack of pinna cues, causes only minor degradation in terms of externalization. This confirms the importance of dynamic cues for externalization, originating from head rotations of the subjects [17], which are well approximated by the

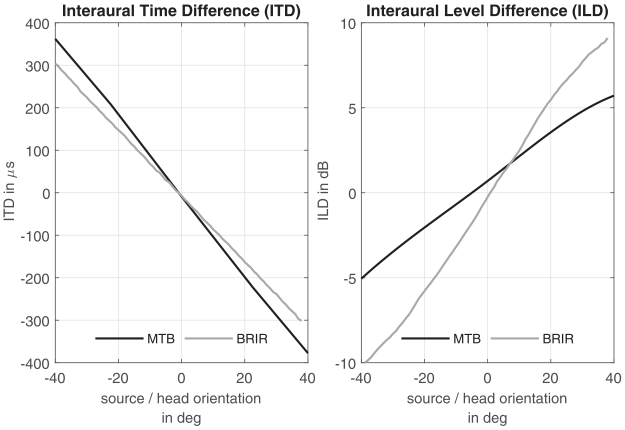


Fig. 9. Interaural time difference (ITD, left) and interaural level difference (ILD, right) for the motion-tracked binaural (MTB, black) and the binaural room impulse response (BRIR)-based reference (gray), measured in the studio for frontal source position.

MTB signals (cf. Fig. 9), even if the spherical array does not exactly match the human head, which is typically less wide than deep [5]. These dynamic cues seem to be effective not only for non-individual binaural synthesis [18], but also for pseudobinaural synthesis without pinna cues.

The fact that sound sources with MTB rendering were perceived slightly closer in distance than the reference, is probably due to a combination of the slight degradation in externalization and the high-frequency spectral boost of the frequency response, which will also affect distance perception. The slight degradation in localizability can most likely be attributed to spectral differences due to missing pinna cues and the slight mismatch between ITD and ILD cues of the MTB (cf. Fig. 9). While ITD differences between true and pseudobinaural signals are below the just noticeable difference (JND) [19, 20], ILD differences clearly exceeded the JND [21, 20].

The overall assessment of these deficits, however, should take into account that in 12 out of 14 conditions these were only observed for the noise stimulus, while for speech, only a slight deviation in tone color remained as an artifact, and no significant degradation in spatial characteristics were observed.

Particularly noteworthy is the fact that there were only minor perceived differences in the source elevation, although pinna cues were traditionally considered the most important for perceiving the elevation of a source [22]. While the spherical head model itself provides only weak cues for the localization of elevated sound sources for *static* sound rendering, our study confirms recent other work showing that motion cues induced by head movements cause changes in the ILD and ITD and can provide missing information to improve the perception of elevation [17, 23]. In our listening test, these spherical head motion cues were sufficient for good localization of source elevation and for an almost unrestricted plausibility of elevated sound sources.

In line with results of a previous study [6], differences between 8 and 16-channel reproduction were small and statistically insignificant, suggesting that a high quality pseudobinaural rendering can already be achieved with 8 microphone capsules.

4 CONCLUSIONS

The current study has investigated the perceptual qualities of pseudobinaural recordings. These signals, which can be recorded with a circular microphone array on a rigid sphere with the dimensions of the human head, include head-related, dynamic cues but no pinna cues. For binaural reproduction, unlike a static dummy head, they offer dynamic cues that provide good externalisation, a largely plausible spatial sound image and a high perceived naturalness of the scene. Even the elevation of sound sources was shown to be partially encoded in the elevation-dependent, dynamic modulation of the ITD and ILD. Unlike a data-based, object-related binaural synthesis, they do not require the binaural impulse responses of each source and receiver to be measured or estimated beforehand. And unlike a BRIR resynthesis based on a sound field analysis using spherical microphone arrays, they require only 8 channels for transmission, while a spherical harmonic (SH) recomposition of the BRIR was shown to require a much higher number of microphones to achieve a better sound quality than the pseudobinaural encoding provided by an MTB microphone [2]. Pseudobinaural recordings can thus provide a conceptually and technically simple yet acoustically convincing approach for an immersive live transmission of complex audio scenes.

5 ACKNOWLEDGMENT

This work was funded by the German Research Fundation (DFG, FOR 1557). The authors would like to thank Steffen Lepa for his advice on the statistical analyses and Georg Neumann GmbH and Sennheiser GmbH & Co. KG for their cooperation in the construction of the MTB Microphone Array.

6 REFERENCES

[1] V. R. Algazi, R. O. Duda, and D. M. Thompson, “Motion-Tracked Binaural Sound,” *J. Audio Eng. Soc.*, vol. 52, no. 11, pp. 1142–1156 (2004 Nov.).

[2] B. Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*, Ph.D. thesis, TU Berlin (2016 Jan.).

[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 Sep.), doi:10.3813/AAA.918562.

[4] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, “A Spatial Audio Quality Inventory (SAQI),” *Acta Acust united Ac*, vol. 100, no. 5, pp. 984–994 (2014 Oct.), doi:10.3813/AAA.918778.

[5] V. R. Algazi, C. Avendano, and R. O. Duda, “Estimation of a Spherical-Head Model from Anthropometry,” *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 472–479 (2001 Jun.).

[6] A. Lindau and S. Roos, “Perceptual Evaluation of Discretization and Interpolation for Motion-Tracked Binaural (MTB) Recordings,” presented at the *26th Tonmeistertagung*, pp. 4087–4096 (2010 Jan.).

[7] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, “Design Criteria for Headphones,” *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 218–232 (1995 Apr.).

[8] F. Brinkmann, A. Lindau, S. Weinzierl, G. Geissler, S. Van De Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, “The FABIAN Head-Related Transfer Function Data Base,” (2017 Feb.), doi:10.14279/depositonce-5718.2.

[9] K. R. Murphy and B. Myors, “Testing the Hypothesis That Treatments Have Negligible Effects: Minimum-Effect Tests in the General Linear Model.” *Journal of Applied Psychology*, vol. 84, no. 2, pp. 234–248 (1999 Apr.), doi:10.1037/0021-9010.84.2.234.

[10] F. Schultz, A. Lindau, M. Makarski, S. Weinzierl, and T. Berlin, “An Extraaural Headphone for Optimized Binaural Reproduction,” presented at the *26th Tonmeistertagung*, pp. 702–714 (2010).

[11] F. Brinkmann, A. Lindau, M. Vrhovnik, and S. Weinzierl, “Assessing the Authenticity of Individual Dynamic Binaural Synthesis,” presented at the *EAA Joint Symposium on Auralization and Ambisonics*, pp. 62–68 (2014 Apr.), doi:10.14279/depositonce-4103.

[12] A. Lindau, S. Weinzierl, and H. Maempel, “FABIAN - An Instrument for Software-Based Measurement of Binaural Room Impulse Responses in Multiple Degrees of Freedom,” presented at the *24th Tonmeistertagung*, pp. 621–625 (2006).

[13] A. Lindau, J. Estrella, and S. Weinzierl, “Individualization of Dynamic Binaural Synthesis by Real Time Manipulation of ITD,” presented at the *128th Convention of the Audio Engineering Society*, (2010 May), convention paper 8088.

[14] S. Ciba, A. Wlodarski, and H.-J. Maempel, “WhisPER– A New Tool for Performing Listening Tests,” presented at the *126th Convention of the Audio Engineering Society* (2009 May), convention paper 7749.

[15] AES Standards Comittee, “AES69-2015: AES Standard for File Exchange - Spatial Acoustic Data File Format” (2015).

[16] J. Carpenter and J. Bithell, “Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians,” *Statistics in Medicine*, vol. 19, no. 9, pp. 1141–1164 (2000 Apr.), doi:10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F.

[17] K. I. McAnally and R. L. Martin, “Sound Localization with Head Movement: Implications for 3-D Audio Displays,” *Frontiers in Neuroscience*, vol. 8, p. 210 (2014 Aug.), doi:10.3389/fnins.2014.00210.

[18] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. Katz, and C. De Boishéraud, “Influence of Head Tracking on the Externalization of Speech Stimuli for Non-Individualized Binaural Synthesis,” *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 2011–2023 (2017 Mar.), doi:10.1121/1.4978612.

[19] J. E. Mossop and J. F. Culling, “Lateralization of Large Interaural Delays,” *J. Acoust. Soc. Am.*, vol. 104, no. 3, pp. 1574–1579 (1998 Sep.), doi:10.1121/1.424369.

[20] S. Klockgether and S. van de Par, “Just Noticeable Differences of Spatial Cues in Echoic and Anechoic Acoustical Environments,” *J. Acoust. Soc. Am.*, vol. 140, no. 4, pp. EL352–EL357 (2016 Oct.), doi:10.1121/1.4964844.

[21] H. Wang, C. Zhang, and Y. Wu, “Just Noticeable Difference of Interaural Level Difference to Frequency and Interaural Level Difference,” presented at the *140th Convention of the Audio Engineering Society* (2016 May), convention paper 9511.

[22] V. R. Algazi, C. Avendano, and R. O. Duda, “Elevation Localization and Head-Related Transfer Function Analysis at Low Frequencies,” *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1110–1122 (2001 Dec.), doi:10.1121/1.1349185.

[23] S. Bögelein, F. Brinkmann, D. Ackermann, and S. Weinzierl, “Localization Cues of a Spherical Head Model,” presented at the *Fortschritte der Akustik: Tagungsband der 44. DAGA*, pp. 347–350 (2018 Mar.).

THE AUTHORS

David Ackermann

Felicitas Fiedler

Fabian Brinkmann

Martin Schneider

Stefan Weinzierl

David Ackermann has received his M.Sc. degree in Audiocommunication and Technology from TU Berlin, Berlin, Germany, in 2015. He has been a research associate in the Audio Communication Group at TU Berlin since 2015, where his research is focused in the field of virtual and musical acoustics. His research includes the investigation of the time-dependent behavior of natural acoustic sound sources and their auralization in virtual acoustic environments.

•

Felicitas Fiedler received her M.Sc. degree in Audio Communication and Technology from TU Berlin, Berlin, Germany, in 2018. In her master's thesis, she dealt with the motion-tracked binaural method for recording spatial sound. Since 2018, she has been working in an acoustic engineering office in Berlin, focusing on room acoustics.

•

Fabian Brinkmann received his M.A. degree (magister artium) in Communication Sciences and Technical Acoustics from TU Berlin, Berlin, Germany. Since 2011, he has been a research associate at the Audio Communication Group from TU Berlin and is associated with the DFG research consortium SEACEN, in which he completed his Ph.D. in the field of signal processing and evaluation approaches for spatial audio.

•

Martin Schneider joined the microphone development department of Georg Neumann GmbH, Berlin, in 1992, after obtaining his diploma in electrical engineering from Technical University Berlin. Since 2008, he has lectured on Electroacoustics at Hochschule für Musik, Detmold, at the Tonmeister institute, and the Microphone Lab at Technical University Berlin, at the Institute for Audio Communication.

•

Stefan Weinzierl is head of the Audio Communication Group at the Technische Universität Berlin. His research is focused on audio technology, virtual acoustics, room acoustics, and musical acoustics. With a diploma in physics and sound engineering, he received his Ph.D. in musical acoustics. He is coordinating a master's program in Audio Communication and Technology at TU Berlin and has coordinated international research consortia in the field of virtual acoustics (SEACEN) and music information retrieval (ABC_DJ).