Acta Acustica 2021, **5**, 30

https://doi.org/10.1051/aacus/2021025

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

SCIENTIFIC ARTICLE OPEN ACCESS

Evaluation of head-tracked binaural auralizations of speech signals generated with a virtual artificial head in anechoic and classroom environments

Mina Fallahi[1], Martin Hansen[1], Simon Doclo[2,4], Steven van de Par[2,4], Dirk Püschel[3], and Matthias Blau[1,4,*]

[1] Institut für Hörtechnik und Audiologie, Jade Hochschule, Ofener Str. 16, 26121 Oldenburg, Germany
[2] Carl von Ossietzky University of Oldenburg, Department of Medical Physics and Acoustics, Carl-von-Ossietzky-Straße 11, 26129 Oldenburg, Germany
[3] Akustik Technologie Göttingen, Bunsenstraße 9c, 37073 Göttingen, Germany
[4] Cluster of Excellence "Hearing4all"

Received 7 April 2021, Accepted 21 June 2021

Abstract – In order to realize binaural auralizations with head tracking, BRIRs of individual listeners are needed for different head orientations. In this contribution, a filter-and-sum beamformer, referred to as virtual artificial head (VAH), was used to synthesize the BRIRs. To this end, room impulse responses were first measured with a VAH, using a planar microphone array with 24 microphones, for one fixed orientation, in an anechoic and a reverberant room. Then, individual spectral weights for 185 orientations of the listener's head were calculated with different parameter sets. Parameters included the number and the direction of the sources considered in the calculation of spectral weights as well as the required minimum mean white noise gain (WNG_m). For both acoustical environments, the quality of the resulting synthesized BRIRs was assessed perceptually in head-tracked auralizations, in direct comparison to real loudspeaker playback in the room. Results showed that both rooms could be auralized with the VAH for speech signals in a perceptually convincing manner, by employing spectral weights calculated with 72 source directions from the horizontal plane. In addition, low resulting WNG_m values should be avoided. Furthermore, in the dynamic binaural auralization with speech signals in this study, individual BRIRs seemed to offer no advantage over non-individual BRIRs, confirming previous results that were obtained with simulated BRIRs.

1 Introduction

In binaural technology, head related transfer functions (HRTFs) play a key role in preserving the spatial attributes of a sound field. In a reverberant environment, binaural room impulse responses (BRIRs) are typically used, which combine the information contained in the HRTFs with the acoustical information of the room. An important application of binaural technology is the auralization of an environment, where a (dry) signal is convolved with measured or modeled BRIRs and presented over headphones [1]. For simulation-based auralizations, the room can be simulated via geometrical acoustic models representing the direct and reflected sound propagation from the source to the listener [2, 3], whereas for measurement-based auralizations, the BRIRs are measured, either with individual listeners or more commonly with so-called artificial heads. Dynamic head-tracked presentation of the auralized environment can greatly enhance the realism of the playback by reducing localization ambiguities and improving the externalization [4–6]. To enable a dynamic auralization, the BRIRs need to be measured for different head orientations [7, 8]. However, this is a very time-consuming task, especially if BRIRs for different head orientations need to be measured individually and in different environments.

In order to avoid repeated BRIR measurements for different head orientations, a few methods have been proposed based on much simpler room impulse response measurements with subsequent modifications. For example, in [9] it was suggested to adapt the auditory scene captured by both microphones of an artificial head to head movements by modifying the cues contained in the binaural signals. In [10], the direct and diffuse parts of the omnidirectional room impulse responses were extracted and modified. The direct and early reflection parts were convolved with the HRTFs of different directions.

Alternatively, microphone arrays have been used to capture the spatial sound field. The captured signals can be then processed to achieve a dynamic head-tracked presentation. For example, with the motion-tracked binaural (MTB) system, a rigid sphere of the size of an average head with microphones distributed on its equator is used to

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

capture the sound. According to the listener's head orientation, the signals captured by the microphones nearest to an ear are interpolated to result in left and right binaural signals [11, 12]. In other approaches, microphone arrays were used together with beamforming methods [13–16], i.e. spectral weights were applied to the captured microphone signals to directly map them to the binaural signals. These approaches offer the advantage of individualizing the recorded signals for individual listeners, by including their individual HRTFs in the binaural signals.

As an alternative to directly mapping the microphone signals to binaural outputs, one can also use an intermediate layer, e.g., in the spherical harmonics domain. In such an approach, the microphone signals are first mapped to the spherical harmonics domain to represent that spatial sound field and subsequently rendered to binaural signals [17–20]. One limitation is the upper frequency limit due to the order truncation and spatial aliasing errors, as the sound field can only be captured by a finite number of microphones. Solutions have been suggested to compensate for this problem [21, 22].

Other approaches employ the spatial decomposition method (SDM) [23, 24] or a parametric time-frequency domain decomposition [25, 26] of B-format microphone signals and assign a Direction of Arrival to those impulse response components that were found to be directional in the decomposition.

In this paper, we consider the beamforming method proposed in [13], where a filter-and-sum beamformer was used to synthesize horizontal HRTFs using a planar microphone array with 24 microphones, referred to as the virtual artificial head (VAH) (see Fig. 1). In [13], the left and right spectral weights were calculated by minimizing a narrow-band least-squares cost function, i.e. minimizing the deviation between the desired and synthesized left and right complex-valued HRTF directivity patterns, respectively, and including regularization to improve robustness [27]. A total of 24 horizontal source directions (15° azimuthal resolution) were considered in the calculation of spectral weights. The synthesized HRTFs were evaluated perceptually in terms of localization, spectral coloration and overall performance, in a static scenario (i.e. without head tracking) for six relevant horizontal directions in an anechoic environment. Out of the six evaluated horizontal directions, three coincided with directions which were also considered in the calculation of the spectral weights. At these directions, the VAH performed as perceptually successful, whereas for the other three directions the results degraded [13]. In [28], the calculation of the spectral weights was modified to include lower and upper constraints on the spectral distortion, such that the deviation in the interaural level difference (ILD) does not exceed 2 dB at any of the synthesis directions, motivated by the reported Just Noticeable Differences in ILD deviations [29]. Simulation results showed that with constraints on spectral distortion, the spatial resolution of the synthesized HRTFs could be improved compared to [13] and [27].

In a previous study [30], this method was also used for a simulated 32-channel microphone array to derive

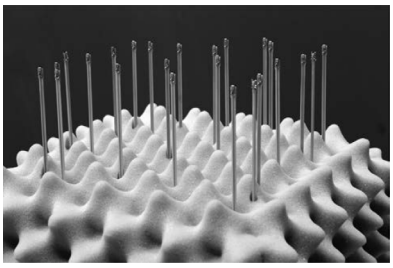

Figure 1. Virtual artificial head (VAH) in this study: planar microphone array with 24 microphones and an extension of 20 cm × 20 cm [13].

synthesized HRTFs for individual listeners. These HRTFs were then used to simulate BRIRs of a lecture room for different head orientations. The room was simulated based on recorded reverberation times and coarse geometries using the RAZR room simulation package [3]. The synthesized BRIRs with the VAH were rated slightly, but not significantly, lower than original BRIRs. The study offered a starting point for evaluating the VAH in dynamic auralizations.

The study in [30] focused on comparing measured BRIRs to various variants of simulated BRIRs (including one featuring the VAH). With simulations as done in [30], optimized solutions for the VAH could not be evaluated with respect to robustness. However, in practice, the robustness of the microphone array is important, because certain solutions will be highly sensitive to small errors in microphone positions and characteristics, as well as to microphone self-noise. Therefore, the present study uses real measurements with the VAH instead of simulations. The same microphone array as in [13] was used together with constraints on spectral distortion as proposed in [28] to synthesize a set of individual BRIRs for 185 different head orientations. Different constraint parameters were considered when using the method in [28], namely the discrete source directions and a parameter related to robustness. The individually synthesized BRIRs as well as measured non-individual BRIRs of a conventional artificial head and a rigid sphere were used for head-tracked auralizations of two acoustically different environments (anechoic and classroom), followed by perceptual evaluations with direct comparison to real loudspeaker signals. The specific research questions to be addressed were: (1) How well does the VAH perform in head-tracked auralizations? (2) Which constraint parameters lead to the best performance of the VAH for auralizing the considered environments? (3) What is the influence of a reverberant environment compared to an anechoic environment on the performance of the VAH? and (4) How well do individually synthesized BRIRs generated with the VAH perform compared to non-individual BRIRs of an artificial head?

The paper continues with a review of the methods and parameters, which were used to calculate the individual spectral weights in Section 2. Section 3 describes the method used for the signal preparation as well as for the perceptual evaluations. Perceptual results for the experiments in the reverberant and anechoic environments are

presented in Sections 4 and 5, respectively, and the results are discussed in Section 6.

2 Virtual artificial head: methods and parameters

In this section, the VAH will be introduced that will be used in the perceptual evaluation in this paper. The VAH is a filter-and-sum beamformer which is optimized using a cost function and certain constraints.

2.1 Calculation of spectral weights using constrained optimization

The virtual artificial head (VAH) as a filter-and-sum beamformer consists of N spatially distributed microphones. The directivity pattern of the VAH at frequency f and direction $\boldsymbol{\Theta} = (\theta, \phi)$, with θ the azimuth angle and ϕ the elevation angle, is given by,

$$H(f, \boldsymbol{\Theta}) = \mathbf{w}^H(f)\mathbf{d}(f, \boldsymbol{\Theta}), \tag{1}$$

with $(\cdot)^H$ indicating the Hermitian transpose. The $N \times 1$ steering vector $\mathbf{d}(f, \boldsymbol{\Theta})$ describes the free-field acoustical transfer function at frequency f between the source at direction $\boldsymbol{\Theta}$ and the N microphones, and the $N \times 1$ vector $\mathbf{w}(f)$ contains the complex-valued spectral weights for each of the N microphones. The aim was to synthesize the desired directivity pattern $D_\zeta(f, \boldsymbol{\Theta}_k)$, $\zeta \in \{L, R\}$, of the left or right HRTFs at P discrete directions $\boldsymbol{\Theta}_k$, $k = 1, 2, \ldots, P$. After transferring the spectral weights $\mathbf{w}_\zeta(f)$, calculated separately for the left and right ears, to filter coefficients of an FIR filter using the inverse Fourier transform, the signals captured with the microphones of the VAH are filtered with these FIR filters and added up to result in binaural left and right signals.

The desired directivity patterns $D_\zeta(f, \boldsymbol{\Theta}_k)$ were spectro-spatially smoothed versions of the original HRTF directivity patterns, using the method presented in [31]. This spectro-spatial smoothing has been shown to be a beneficial step that allowed to obtain a generally better approximation of the desired directivity pattern, without introducing perceptible degradations [31]. The spectral weights $\mathbf{w}_L(f)$ and $\mathbf{w}_R(f)$ were calculated by minimizing a narrow-band least-squares cost function, which is defined as the sum of the squared absolute differences between desired and synthesized directivity patterns over P discrete directions, i.e.,

$$J_{\mathrm{LS}}(\mathbf{w}_\zeta(f)) = \sum_{k=1}^{P} |H_\zeta(f, \boldsymbol{\Theta}_k) - D_\zeta(f, \boldsymbol{\Theta}_k)|^2, \tag{2}$$

where H_ζ indicates the synthesized HRTFs for the left and right ears as defined in Equation (1). The cost function J_{LS} was minimized separately for the left and the right ears. Aiming at achieving a small synthesis error at all P directions, it was proposed in [28] to impose constraints onto the spectral distortion (SD), defined as,

$$\mathrm{SD}_\zeta(f, \boldsymbol{\Theta}_k) = 10 \lg \frac{|\mathbf{w}_\zeta^H(f)\mathbf{d}(f, \boldsymbol{\Theta}_k)|^2}{|D_\zeta(f, \boldsymbol{\Theta}_k)|^2} \mathrm{dB}. \tag{3}$$

Constraints were imposed on the SD such that at each direction $\boldsymbol{\Theta}_k$:

$$L_{\mathrm{Low}} \leq \mathrm{SD}_\zeta(f, \boldsymbol{\Theta}_k) \leq L_{\mathrm{Up}}, \quad k = 1, 2, \ldots, P, \tag{4}$$

where L_{Up} and L_{Low} denote the upper and lower boundary, respectively. An additional constraint was imposed onto the *mean* white noise gain (WNG_m) [27], which is defined as the ratio between the mean output power of the microphone array over all P directions and the output power of spatially uncorrelated white noise, i.e.,

$$\mathrm{WNG_m} = 10 \lg \left(\frac{1}{P} \sum_{k=1}^{P} \frac{|\mathbf{w}_\zeta^H(f)\mathbf{d}(f, \boldsymbol{\Theta}_k)|^2}{\mathbf{w}_\zeta^H(f)\mathbf{w}_\zeta(f)} \right) \mathrm{dB} \geq \beta, \tag{5}$$

where β denotes the minimum desired WNG_m in dB. This additional constraint was applied in order to increase the robustness of the VAH against small deviations in microphone positions and characteristics and limit microphone self-noise amplification. To solve the constrained optimization problem of minimizing J_{LS} in Equation (2) subject to $P + 1$ constraints defined in Equations (4) and (5), an iterative Interior-Point algorithm, as implemented in function fmincon in the MATLAB optimization toolbox (ver. R2018b) was used.

2.2 Microphone array and constraint parameters

In this study, the planar microphone array shown in Figure 1 was used. This microphone array consisted of 24 sensors, each consisting of two MEMS microphones (Analog Devices ADMP 504 Ultralow Noise Microphone), with a Golomb-based topology. This microphone array was previously evaluated perceptually in [13] for a static scenario (i.e. without head tracking).

The upper and lower boundaries L_{Up} and L_{Low} for the SD constraints in Equation (4) were chosen as 0.5 dB and −1.5 dB, respectively. Satisfying the SD constraints with these values of L_{Up} and L_{Low} results in a maximum deviation of 2 dB in the resulting interaural level differences (ILDs) at all P directions. A deviation of 2 dB was considered reasonable based on the reported Just Noticeable Differences in ILD deviations [29]. For the lower boundary of the WNG_m, i.e. β in Equation (5), two values of 0 dB and −10 dB were considered, labeled as β_0 and β_{-10} in the remaining discussion. The choice of $\beta = 0$ dB was based on the results in [32], while $\beta = -10$ dB was chosen to investigate the effect of a lower resulting WNG_m and a reduced robustness.

It should be noted that the P directions considered in the calculation of the spectral weights, i.e. both in the cost function in Equation (2) as well as in the SD constraints in Equation (4), have a major influence on the resulting synthesized HRTFs, spectral distortion and WNG_m. It is therefore interesting to investigate the extent to which it is necessary to include directions other than horizontal directions into the calculation of the spectral weights, in order to account for non-horizontal source positions as well as room reflections. Three cases for P were considered in the study: (1) $P = 72$ horizontal directions (5° azimuthal resolution),

(2) $P = 3 \times 72 = 216$ directions from elevations −15°, 0° and +15° and (3) $P = 3 \times 72 = 216$ directions from elevations −30°, 0° and +30°, labeled as **V0, V0 ± 15** and **V0 ± 30**, respectively, in the remaining discussion. Table 1 summarizes the constraint parameters P and β used for the calculation of the spectral weights in this study.

As an example, spectral weights were calculated with a set of measured steering vectors and the individual HRTFs of one of the subjects in this study (subject 1) for different values of P and β. The calculated spectral weights were then applied to the same measured steering vectors using Equation (1) to result in the synthesized HRTFs for subject 1 for the frontal head orientation. Figure 2 shows the resulting SD for the synthesized left HRTFs at elevations 0°, 15° and 22.5°, as well as the resulting WNG_m. The two upper parts (Figs. 2a and 2b) show the results for $\mathrm{V0}/\beta_0$ and $\mathrm{V0}/\beta_{-10}$, respectively. At elevation 0°, it can be observed that up to about 5 kHz, the SD constraints as well as WNG_m constraints could be satisfied. However, at frequencies above 5 kHz, the SD constraints could not always be satisfied. Above 5 kHz, the WNG_m constraints could be satisfied for all frequencies with $\mathrm{V0}/\beta_{-10}$, whereas with $\mathrm{V0}/\beta_0$, it was only the case above 8 kHz. At elevations 15° and 22.5°, the resulting SD clearly increased compared to the resulting SD at elevation 0°, since these non-horizontal directions were not included in the calculation of the spectral weights.

The two lower parts (Figs. 2c and 2d) show the results for $\mathrm{V0} \pm 15/\beta_0$ and $\mathrm{V0} \pm 15/\beta_{-10}$, respectively. Compared to the results shown in Figures 2a and 2b, for frequencies up to about 4 kHz, the resulting SD at elevation 15° clearly improved. The inclusion of non-horizontal directions also slightly improved the resulting SD at elevation 22.5°, although directions from this elevation were not included in the constrained optimization. At the same time, the resulting SD at elevation 0° deteriorated. In addition, the WNG_m constraint could, with a few exceptions, not be satisfied for frequencies below 4 kHz.

2.3 Spectral weights for head-tracked binaural renderings

An important feature of the VAH is the possibility of calculating the spectral weights for different head orientations without requiring individual HRTFs for these head orientations. This enables head rotations to be easily taken into account via head tracking during signal playback. For a given head orientation $\Theta_h = (\theta_h, \phi_h)$, with θ_h and ϕ_h denoting the horizontal and vertical angles of the head orientation, respectively, spectral weights can be calculated by considering the desired directivity pattern $D_\zeta(f, \Theta_k)$, $k = 1, 2, \ldots, P$, together with spatially shifted steering vectors $\mathbf{d}(f, \Theta_s)$ with $\Theta_s = (\theta_k + \theta_h, \phi_k + \phi_h)$ in Equations (1)–(5). This can be interpreted as a virtual rotation of the VAH to head orientation Θ_h.

To enable a head-tracked binaural signal playback with the VAH in this study, spectral weights were a priori calculated for each of the six parameter sets listed in Table 1 for $37 \times 5 = 185$ head orientations (37 horizontal directions $-90° \leq \theta_h \leq 90°$ in 5° steps and 5 vertical directions $-15° \leq \phi_h \leq 15°$ in 7.5° steps).

Table 1. Overview of values chosen for the parameters P and β, resulting in 6 sets of spectral weights. Each set of spectral weights was calculated for 185 head orientations.

Label	Constraint parameters P and β
$\mathrm{V0}/\beta_0$	$P = 72$ (elevation: 0°) $\beta = 0$ dB
$\mathrm{V0} \pm 15/\beta_0$	$P = 216$ (elevations: −15°, 0°, 15°) $\beta = 0$ dB
$\mathrm{V0} \pm 30/\beta_0$	$P = 216$ (elevations: −30°, 0°, 30°) $\beta = 0$ dB
$\mathrm{V0}/\beta_{-10}$	$P = 72$ (elevation: 0°) $\beta = -10$ dB
$\mathrm{V0} \pm 15/\beta_{-10}$	$P = 216$ (elevations: −15°, 0°, 15°) $\beta = -10$ dB
$\mathrm{V0} \pm 30/\beta_{-10}$	$P = 216$ (elevations: −30°, 0°, 30°) $\beta = -10$ dB

The spatial resolution of 5° for the horizontal head movements was chosen in accordance with the resolution reported to be sufficient for non-critical signals such as music [33]. It should be mentioned that the measured steering vectors were not available for all shifted directions Θ_s in the vertical direction (see Sect. 3.1). For such cases, the steering vectors were shifted to the nearest available vertical direction.

3 Methods

This study consisted of two experiments with measurements and perceptual evaluations in two different acoustical environments: a reverberant lecture room (Experiment 1) and an anechoic room (Experiment 2). Experiment 2 was performed after completing Experiment 1 and was motivated by questions arisen from the results of Experiment 1. The methods and technical implementations were to a large extent the same for both experiments. The information provided in this section applies to both experiments. Specific information on each experiment (room characteristics, source and listener positions) are provided in more detail in Sections 4 and 5 as well as in Figure 3. The description of the applied methods starts with introducing the preparatory measurements in Section 3.1, followed by the methods applied for the acquisition of BRIRs in Section 3.2. Technical implementation for listening tests and the criterion to exclude non-consistent ratings are discussed in Sections 3.3 and 3.4, respectively.

3.1 Preparatory measurements

Individual head related impulse responses (HRIRs) and VAH steering vectors, $\mathbf{d}$, were measured in an acoustic laboratory (10 m × 7.75 m × 3 m, reverberation time: 0.46 s), using a loudspeaker arc of 1.25 m radius with small active loudspeakers (Speedlink SL-8902-GY Xilu) covering 12 elevations (−30° to +30° in 7.5° steps, 45°,

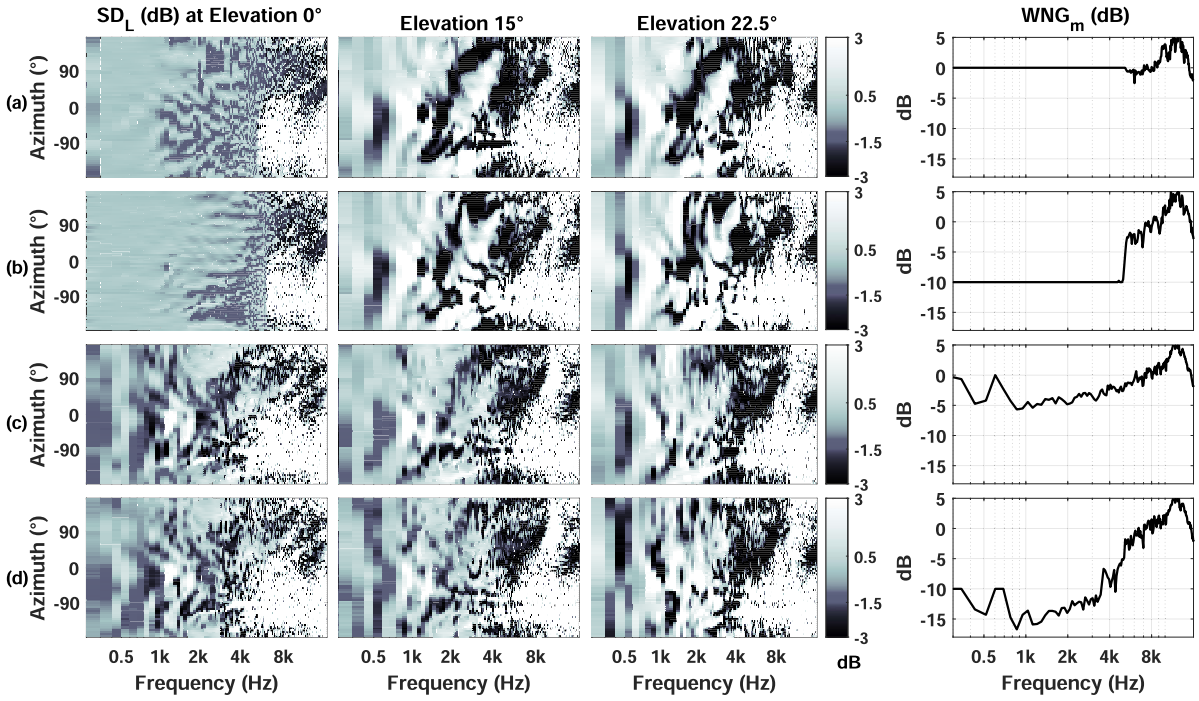

Figure 2. The resulting SD at elevations 0°, 15° and 22.5° (the first three columns from left) and the resulting WNG_m (the right column). The spectral weights were calculated with (a): $\mathrm{V0}/\beta_0$, (b): $\mathrm{V0}/\beta_{-10}$ (c): $\mathrm{V0} \pm 15/\beta_0$, (d): $\mathrm{V0} \pm 15/\beta_{-10}$. The results are shown for the left ear of subject 1 in this study.

60° and 75°), and a 32-channel audio interface (Antelope Orion). For measuring individual HRIRs, subjects were seated with their head positioned in the center of the arc. Similarly, for measuring steering vectors, the VAH was positioned in the center of the arc. The loudspeaker arc was rotated via a turn table around the subject or the VAH in 5° steps. At each azimuthal position of the arc, impulse responses were measured at $f_s = 44\,100$ Hz using the Multiple Exponential Sweep Method (MESM) [34] with modifications as proposed in [35]. The excitation sweeps were 17 s long with 0.35 s shift between subsequent sweeps and covered the frequency range from 100 Hz to $f_s/2$. For the HRIR measurement, subjects wore two MEMS microphones (Knowles PSV0840LR5H) at the entrance of the blocked ear canals, using 3D-printed supports fitted into foam earplugs, see [36] for details. Room reflections were damped by around 15 dB using absorbent foams mounted to the floor and the ceiling. In addition, the measured impulse responses were truncated to 256 samples using a 50-point half-Hann window in order to further eliminate room reflections.

Subsequent to the HRIR measurement and before removing the microphones from the blocked ear canals, individual headphone impulse responses (HPIRs) were measured. The HPIR measurement was repeated nine times, each after repositioning the headphones (Sennheiser HD 800). The pair of HPIRs resulting in the smallest dips in the frequency range between 8 kHz and 12 kHz was chosen for the calculation of the individual inverse HPIRs, as described in [13, 30]. For this inversion, the regularized inversion method in [37] was applied, with the regularization parameter $\beta_{\mathrm{inversion}} = 10$ times the average mean square value of the headphone impulse responses. The individual inverse HPIRs were truncated to a length of 2048 samples.

After transferring the measured HRIRs into the frequency domain, the HRTFs were spectro-spatially smoothed according to [31]. The smoothed desired directivity patterns D and the measured, Fourier-transformed steering vectors were used to calculate the individual spectral weights with $L_{\mathrm{Up}} = 0.5$ dB, $L_{\mathrm{Low}} = -1.5$ dB and the listed values for the parameters P and β in Table 1. The individual spectral weights were calculated for the 185 considered head orientations, as described in Section 2.3.

3.2 BRIR acquisition

In both environments, one listener position and different source positions were defined, which are shown in Figures 3a and 3b. The VAH was placed at the listener position and room impulse responses were measured between the different source positions and the 24 microphones of the VAH. These room impulse responses were filtered with the FIR filters corresponding to the individually calculated left and right spectral weights (for each of the 185 head orientations) and added up over the N channels into the left and right BRIRs. These BRIRs are referred to as **VAH BRIRs**.

In both environments, BRIRs were also measured with a commercial artificial head (KEMAR type 45BB, GRAS Sound & Vibration A/S, Holte, Denmark) as well as with a head-sized rigid sphere (radius = 8.5 cm) with two MEMS

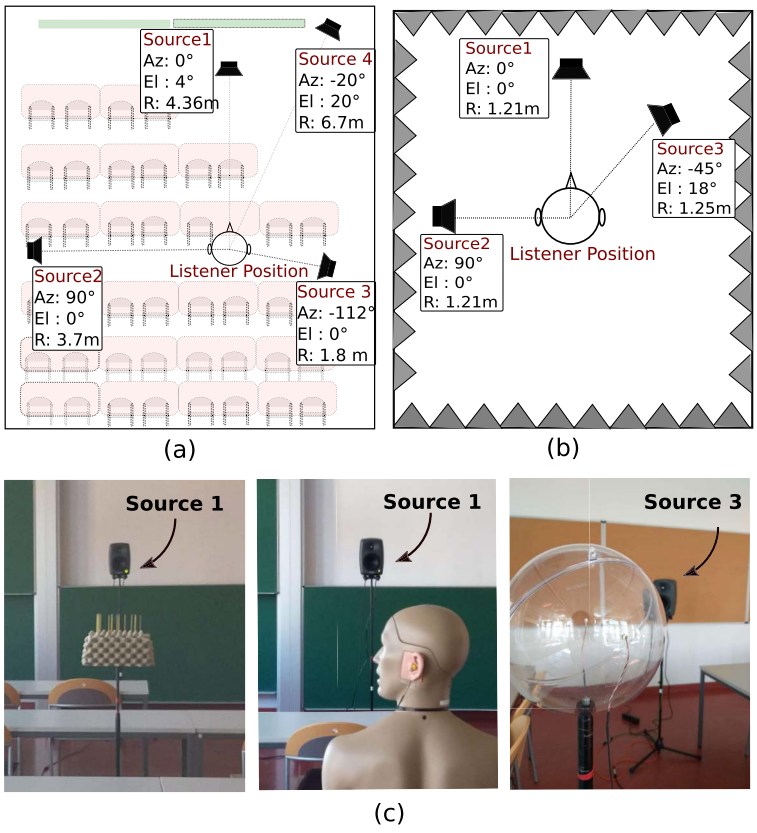

Figure 3. Listener and source positions (Az: azimuth, El: elevation, R: distance to listener) in (a) lecture room (Experiment 1) and (b) anechoic room (Experiment 2). (c): (from left to right) VAH, KEMAR artificial head and the rigid sphere in the lecture room.

Knowles PSV0840LR5H microphones positioned at ±100° on the equator (see Fig. 3c). In order to enable a head-tracked signal presentation with the binaural signals captured with the KEMAR artificial head or the rigid sphere at least for horizontal head orientations, the BRIR measurement was repeated 37 times for 37 horizontal orientations of the artificial head and rigid sphere (−90° to 90° in 5° steps). Note that this scenario of a moving artificial head is obviously non-realistic and cannot be employed in standard applications. It was considered in this study nonetheless, since the usual static scenario for the KEMAR artificial head and the rigid sphere would have been too easy to discriminate from the head-tracked VAH BRIRs in listening tests. The BRIRs measured for different orientations of the KEMAR artificial head or the rigid sphere are referred to as **HTK BRIRs** (**H**ead-**T**racked **K**EMAR) and **HTS BRIRs** (**H**ead-**T**acked **S**phere), respectively.

All BRIRs were measured at $f_s = 44\,100$ Hz, using the MESM method with sweeps of 20 s duration, from 20 Hz to $f_s/2$ with 4 s shift between subsequent excitations. In the lecture room, the measured impulse responses were truncated to a length of 18 000 samples using a 50-point half-Hann window. After convolution with the individual inverse HPIRs, the VAH, HTK and HTS BRIRs were truncated again to a final length of 18 000 samples, corresponding to 408 ms at $f_s = 44\,100$ Hz and a decay of over 40 dB, which enabled to cover the usable dynamic range in the room (see Sect. 4). In the anechoic room, the measured impulse responses were truncated to a length of 1024 samples using a 50-point half-Hann window. After convolution with the individual inverse HPIRs, the VAH, HTK and HTS BRIRs had a final length of 3071 samples.

It should be noted that although the anechoic room could be considered as a free-field environment, the measured or synthesized binaural impulse responses in Experiment 2 are denoted as BRIRs (instead of HRIRs) to reflect the influence of the experimental apparatus in the room.

3.3 Listening test – Technical implementation

To evaluate the quality of the (individually synthesized) VAH BRIRs as well as the (non-individual) HTK and HTS BRIRs, two listening tests (Experiment 1 and Experiment 2) were performed. In both tests, head tracking was employed. During the listening tests, subjects sat at the same listener position as defined for the BRIR measurements. They were asked to rate different binaural presentations with headphones, generated either with VAH BRIRs for different parameter sets or with HTK and HTS BRIRs, in comparison to a reference signal (real loudspeaker playback in the room). Subjects could decide by themselves when to listen to headphone or the reference signal and were asked to take off the headphones when listening to the loudspeaker.

Playback was conveniently switched between the loudspeaker and headphone presentation via a control-switch on the headphones, as implemented in [30]. Subjects had no information about the BRIR condition which was presented at any time. Loudspeakers and their positions, as well as all other features such as visual cues or the arrangement of the objects in the room, remained the same as during the BRIR measurements.

A custom-made head tracker was mounted on the top of the headphones (the same headphones as used for measuring the individual HPIRs) and the real-time head-tracked binaural playback was generated with a custom C++ program based on [38]. The latency caused by the system was 5.8 ms and head tracker data were updated each 10 ms. In both experiments, the signals were played back over an external audio interface (RME Fireface UC). For the headphone signals, a headphone amplifier (Lake People Phone-Amp G103) was used. The loudness of the real sources was adjusted manually by the experimenters to have the same loudness impression as the headphone signals. The different BRIRs were compared with the reference signal in terms of perceptual attributes (the same as used in [30]) "Halligkeit" (Reverberance), "Quellbreite" (Source Width), "Quelldistanz" (Source Distance), "Schallquellenrichtung" (Source Direction) and "Gesamtqualität" (Overall Quality). The perceptual attributes were presented always in the same order as given above, i.e. Experiment 1 started for all subjects with evaluating the attribute Reverberance, continuing to Source Width and so on. The attribute Reverberance was not evaluated in Experiment 2. Therefore, Experiment 2 started with the attribute Source Width and so forth. In order to limit the number of evaluations, the perceptual attribute spectral coloration was not explicitly evaluated but was assumed to be included in the perceptual attribute Overall Quality. To give their ratings with respect to the perceptual attribute Overall Quality, subjects were instructed to exclude all aspects related to the previous attributes and to focus on everything not included yet. Subjects rated the attributes on a 9-point scale with five German labels "schlecht" (bad), "dürftig" (poor), "ordentlich" (fair), "gut" (good) and "ausgezeichnet" (excellent) and four unlabeled intermediate points (the scale point names and their English translations were taken from [39]). To obtain the ratings, a graphical user interface (GUI) was presented to the subjects, with sliders which could be moved with a mouse. Before starting the experiment, subjects could get familiar with the environment, with the GUI as well as with the equipment. The main experiment began after this familiarization by explaining the first perceptual attribute. After completing the ratings for one perceptual attribute and before continuing to the next one, subjects were provided with the explanation of the next perceptual attribute. Perceptual attributes were explained with a short description in German language.

Each of the source positions shown in Figure 3 appeared three times during the evaluation in a randomized order. Subjects were allowed to switch freely between different headphone signals and between headphone and loudspeaker presentations. They were informed that head rotations were permitted in the horizontal and vertical range of ±90° and ±15°, respectively. However, no explicit instruction was given to the subjects to rotate their heads while listening to the signals. Subjects were asked to reset the head tracker by keeping the head to the front and clicking a 'reset' button on the GUI, before evaluating a given source position and perceptual attribute.

As in [30], the stimulus was a dry recorded speech utterance of 15 s duration ("Nordwind und Sonne", text version from the IPA Handbook [40], first sentence), spoken by a female speaker. This audio sample was repeated to a total length of about three minutes to provide the subjects with enough time to compare and rate the different signal presentations. In case that subjects were not finished by the end of the 3-min long signal playback, they could easily repeat the playback from the beginning. For a given source position and perceptual attribute, it took the subjects on average 2.5 min to complete the comparison between different headphone presentations and the reference signal.

Ten normal-hearing subjects (six male, four female, aged 20–52 years old, all having a hearing threshold of 15 dB HL or better verified by a pure tone audiometry between 125 Hz and 8 kHz) participated in the experiments. Eight subjects reported to have extensive experience with perceptual listening tests, while two subjects reported to not have much prior experience. For all subjects, individually measured HRIRs and HPIRs as well as individually calculated spectral weights for parameter sets listed in Table 1 for 185 head orientations were prepared.

It should be mentioned that although no explicit head movement instructions were given, all subjects moved the head during headphone signal presentation. The amplitude and trajectory of head movements varied among the subjects. The intra-subject amplitude of head movements, on the other hand, remained stable across the perceptual attributes. Figure 4 shows exemplary horizontal head orientations of the subjects, collected by the tracker device when listening to headphone presentations of Source 2 and evaluating the perceptual attribute Overall Quality in Expriment 1.

3.4 Exclusion of non-consistent ratings

As already mentioned in Section 3.3, for each perceptual attribute each source position in the room was presented and evaluated three times. To assess the consistency of the ratings over the three repetitions, the Pearson correlation coefficients between the three presentation pairs (1–2, 1–3, 2–3) were calculated separately for each attribute and for each subject. As a measure of repeatability, the mean correlation coefficient ($\bar{r}$) was evaluated according to:

$$\bar{r} = \frac{\alpha}{n + (1 - n)\alpha}, \tag{6}$$

with α indicating the Cronbach's standardized coefficient [41] and n the number of repetitions. With $\alpha > 0.8$ considered as "good" and with $n = 3$ repetitions, ratings with $\bar{r} > 0.57$ were considered as consistent and repeatable (the same values as also used in [30]). If not, the ratings

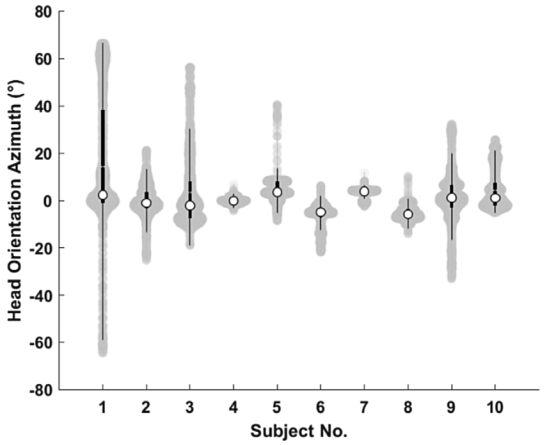

Figure 4. Violin plots showing subjects' horizontal head orientations when listening to headphone presentations of Source 2 and evaluating the perceptual attribute Overall Quality in Expriment 1.

of this subject for the investigated perceptual attribute were excluded. The mean Pearson coefficients $\overline{r}$ for all subjects and for all perceptual attributes in both experiments are shown in Figure 5. For the subjects fulfilling the repeatability criterion, it was supposed that there are no differences between the three presentations. Therefore, the ratings were averaged over three repetitions for further analysis. Please note that the consistency was assumed to vary not only among subjects but also among the perceptual attributes, meaning that the consistent evaluation of some perceptual attributes could have been more challenging than the others. Therefore, it was decided to exclude only the inconsistent ratings instead of completely excluding subjects with partly inconsistent ratings. For the convenience of the reader, the raw data of both listening tests is available at http://doi.org/10.5281/zenodo.4616259.

4 Experiment 1

The first experiment was performed in a lecture room (7.12 m × 11.94 m × 2.98 m) with an average reverberation time of 0.58 s and with six rows of tables and chairs (see Fig. 3a). The listener position was chosen in the third row in the middle slightly shifted to the right, at 1.30 m height, which was assumed to be the height of the ear axis for subjects sitting at the listener position. Four sources were considered in the room: Source 1 (Genelec type 8030c) was located ahead of the listener at a slightly higher position than the ears. Two other sources, Source 2 and Source 3 (Genelec type 8030b) were located at the left and behind the listener at the right side, both at the same height as the ears. Source 4 (Event active studio monitor 20/20 bas V3), was located at the frontal upper right corner of the room at an elevation of about 20°. The sound pressure level at listener position was 60 dBA with signals played back from Source 1 and the background noise in the room was measured at around 20 to 25 dBA, depending on outdoor conditions.

For each source position, the six individually calculated VAH BRIRs, synthesized using the parameter sets listed in Table 1, as well as the non-individual HTK and HTS BRIRs, measured for the KEMAR artificial head and the rigid sphere, respectively, were evaluated.

4.1 Experiment 1: Results

By excluding the non-consistent ratings as described in Section 3.4, the number of subjects was reduced to eight for the perceptual attributes Source Width, Source Distance, Source Direction and Overall Quality (Fig. 5). For the perceptual attribute Reverberance, one subject was excluded. It should be noted that seven of the nine exclusions pertained to the two subjects with less experience (subject 9 and subject 10).

Figure 6 shows the histogram of rating differences between the three presentation pairs after excluding the non-consistent ratings. Between 35% and 48% of the ratings were identical (i.e. the difference was zero), and between 74% and 85% were within ±1 scale units. The symmetrical distribution of differences with respect to zero difference, similarly for all presentation pairs and attributes, indicates that there were no substantial learning effects over time.

Figure 7 shows the perceptual evaluations for the five perceptual attributes, four source positions and eight different BRIR sets. For almost all perceptual attributes and source positions, the VAH BRIRs with V0/β_0 and the HTK and HTS BRIRs were rated similarly high, with median values between good and excellent. In comparison, the VAH BRIRs including non-horizontal directions (V0 ± 15 and V0 ± 30) were rated lower, regardless of the parameter β. Even for Source 4, which was located markedly out of the horizontal plane, the VAH BRIRs with V0 ± 15 and V0 ± 30 were rated lower than the VAH BRIRs optimized using only horizontal directions (V0/β_0 and V0/β_{-10}). For all source positions and perceptual attributes, the VAH BRIRs with V0/β_{-10} were rated lower than the VAH BRIRs with V0/β_0, but higher than the VAH BRIRs with V0 ± 15 and V0 ± 30, regardless of the parameter β.

The Shapiro-Wilk test of normality, applied to the ratings for each combination of source position, BRIR set and perceptual attribute, revealed that the ratings cannot be assumed to be normally distributed for all cases ($p < 0.05$). Therefore, a non-parametric method (Friedman test) was used to statistically analyse the ratings. According to the Friedman test, for 20 out of 40 combinations of BRIR sets and perceptual attributes, a significant effect of the source position could be observed. p-values are shown in Table 2, with bold cases indicating $p < 0.05$. The effect of source position was in 17 of 20 cases significant for the three perceptual attributes Reverberance, Source Width and Source Distance. However, since for each of the evaluated source positions the experiment design focused on the comparison of different BRIR sets, the ratings were averaged over the four source positions in order to statistically

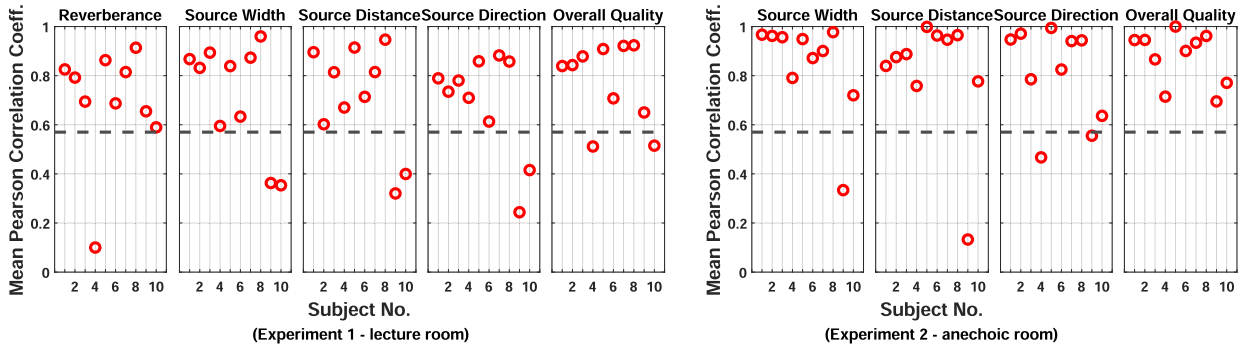

Figure 5. Mean Pearson correlation coefficients $\overline{r}$ for the three presentation pairs (1–2, 1–3, 2–3) as a measure for consistent ratings. The dashed horizontal line indicates the chosen lower threshold for $\overline{r}$. Subjects with a $\overline{r}$ below this threshold were excluded from the evaluations. To calculate the correlations coefficients, 32 combinations (4 loudspeaker positions × 8 BRIR sets) for Experiment 1 and 18 combinations (3 loudspeaker positions × 6 BRIR sets) for Experiment 2 were considered for each of the three presentation pairs.

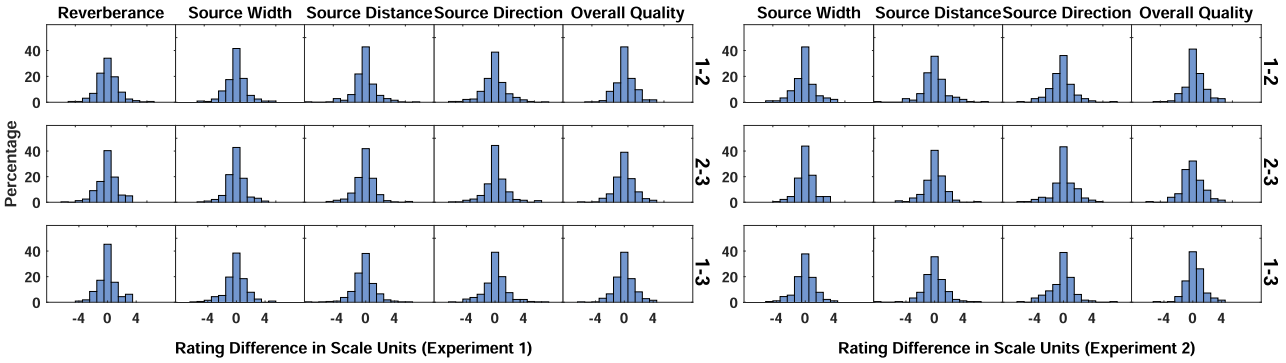

Figure 6. Histogram of rating differences between the three presentation pairs (1–2, 1–3, 2–3) after excluding the non-consistent ratings. Two scale units correspond to the difference between adjacent labeled scale points.

analyse the effect of the BRIR sets. The averaged ratings are shown in Figure 8. As determined by Shapiro-Wilk test of normality, also the ratings averaged over the source positions could not be assumed for all BRIRs to be normally distributed. Therefore, the Friedman test was applied which revealed for all attributes a significant effect of BRIR set ($p < 10^{-4}$). As indicated by multiple comparisons after Friedman test (function friedmanmc in R [42]), significantly lower ratings were given to VAH BRIRs with V0 ± 15/$\beta_{0,-10}$ and V0 ± 30/$\beta_{0,-10}$. For all perceptual attributes, there were no significant differences between the VAH BRIRs with V0/$\beta_{0,-10}$ and the HTK or HTS BRIRs. There were also no significant differences between V0/β_0 and V0/β_{-10}.

5 Experiment 2

The results in Experiment 1 revealed a perceptually successful performance of the VAH BRIRs as well as HTK and HTS BRIRs. The extent to which the room effects might have had an impact on the perception of different BRIRs was however not clear. Reverberation is expected to reduce source localization accuracy by itself, which may interact with the ratings of the subjects. It was interesting to see whether a similar performance with the tested BRIRs can also be achieved in the absence of room effects. Therefore, a similar experiment (Experiment 2) was performed in an anechoic environment. Since in Experiment 1, the ratings for the VAH BRIRs including non-horizontal directions (V0 ± 15/$\beta_{0,-10}$ and V0 ± 30/$\beta_{0,-10}$) were similarly low, the VAH BRIRs with V0 ± 30/β_0 and V0 ± 30/β_{-10} were excluded in Experiment 2.

Experiment 2 was performed in the anechoic room of the *Institut für Hörtechnik und Audiologie* at the Jade University of Applied Sciences in Oldenburg (3.1 m × 3.4 m × 2 m, cutoff frequency 200 Hz). The listener position was chosen in the middle of the room (see Fig. 3b). Three sources (Fostex 6301B) were positioned in the room. Source 1 and Source 2 were located in front and at the left of the listener, respectively, both at the same height as the ears. Source 3 was located at 45° at the right side at an elevation of about 18°. Source 1 and Source 2 in Experiment 2 were considered equivalent to Source 1 and Source 2 in Experiment 1. However, due to practical reasons, Source 3 in Experiment 2, which was chosen to represent the sound source outside the horizontal plane, had a different position than its equivalent (Source 4) in Experiment 1.

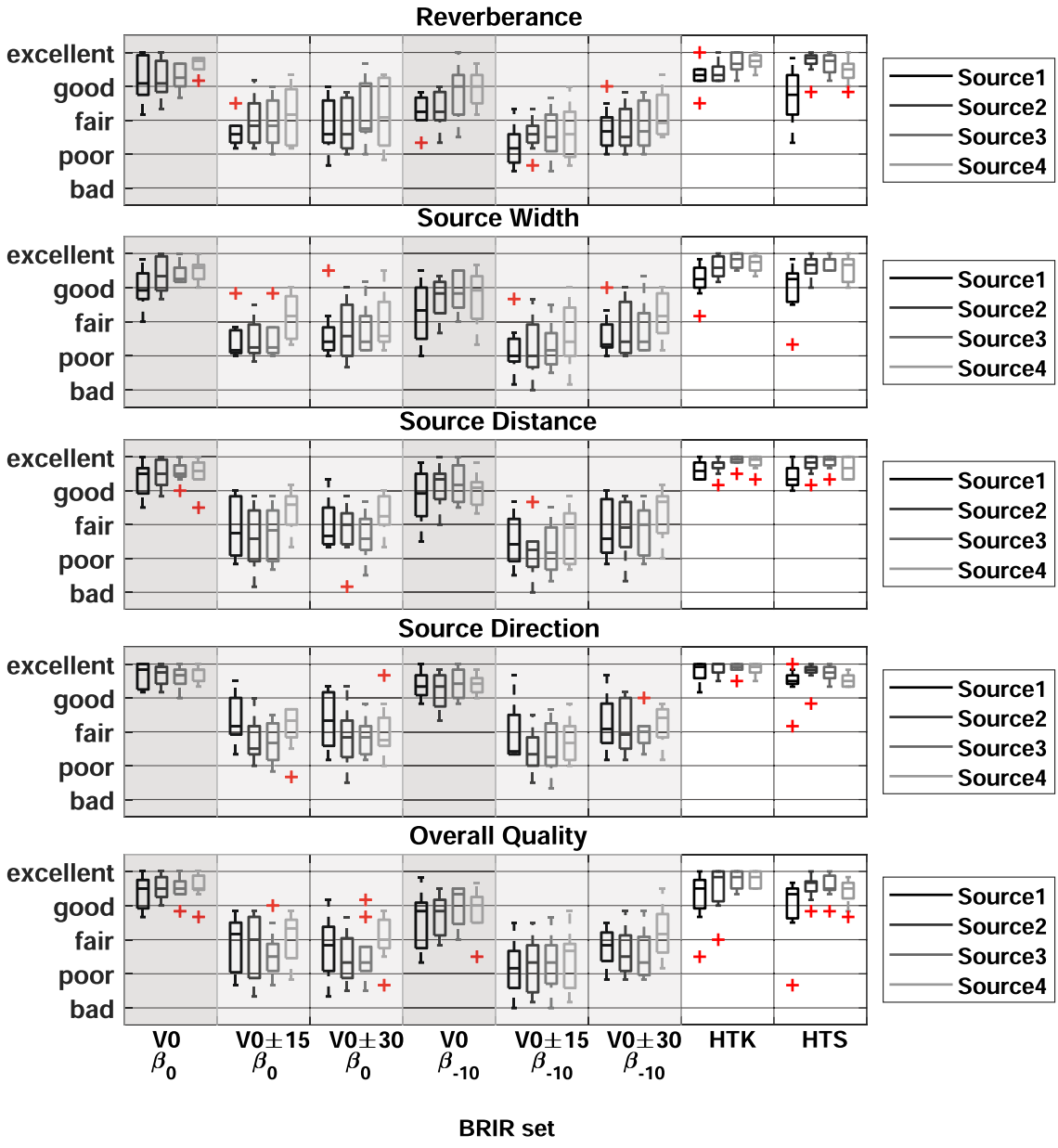

Figure 7. (Experiment 1) Perceptual ratings averaged over three repetitions, for five perceptual attributes, four source positions, and eight different BRIR sets.

Table 2. p-values (Friedman test) for investigating the effect of source position on the ratings given to different BRIR sets for each perceptual attribute in Experiment 1 (Exp. 1) and Experiment 2 (Exp. 2). p-values indicating significant different ratings ($p < 0.05$) are depicted as bold numbers.

BRIR set	Reverberance		Source width		Source distance		Source direction		Overall quality	
	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2
V0/β_0	0.058	–	**0.034**	**0.015**	0.526	0.748	0.666	0.145	0.231	0.581
V0 ± 15/β_0	0.228	–	**0.001**	0.670	**0.010**	**0.020**	**0.015**	**0.011**	0.637	0.575
V0 ± 30/β_0	**0.032**	–	0.286	–	**0.017**	–	0.409	–	0.050	–
V0/β_{-10}	**0.004**	–	0.073	0.428	0.365	0.176	0.436	0.067	0.057	0.478
V0 ± 15/β_{-10}	**0.022**	–	0.190	**0.023**	**0.038**	0.648	0.075	0.115	0.078	**0.011**
V0 ± 30/β_{-10}	**0.006**	–	**0.013**	–	**0.034**	–	0.830	–	**0.011**	–
HTK	**0.012**	–	**0.000**	0.434	**0.021**	0.314	0.625	0.050	0.138	0.335
HTS	**0.002**	–	**0.002**	0.393	**0.003**	0.661	0.060	0.184	**0.020**	0.823

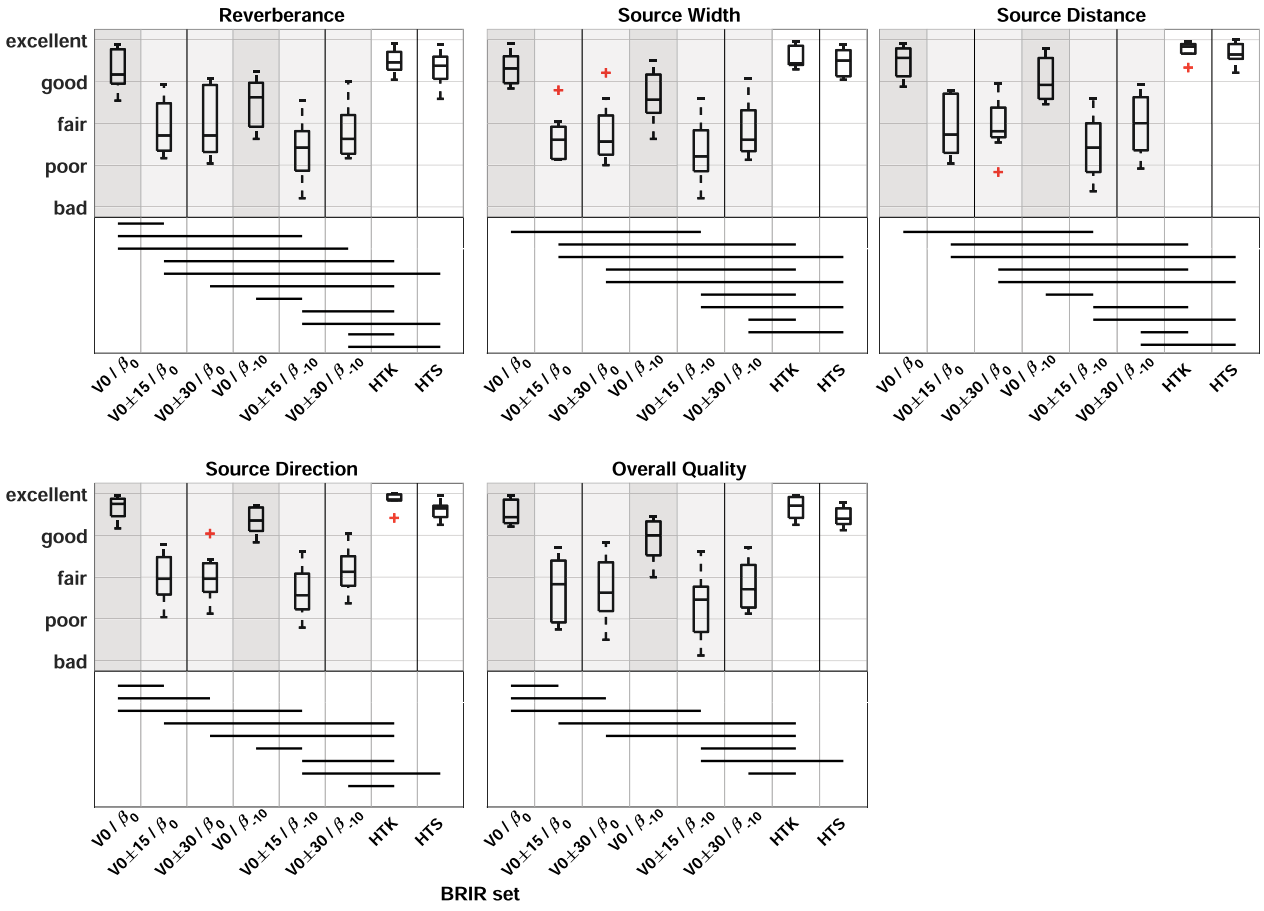

Figure 8. (Experiment 1) Averaged ratings over four source positions for different BRIR sets and perceptual attributes. Significant different ratings are marked with horizontal lines ($p < 0.05$).

Nevertheless, the two non-horizontal sources had similar elevations (20° in Experiment 1 and 18° in Experiment 2). In addition, the azimuthal position of the non-horizontal sources, both in front of the listener on the right side, coincided with one of the azimuthal directions included in the calculation of the spectral weights (5° azimuthal resolution). Consequently, the impact of the constraint parameters chosen for the calculation of the VAH spectral weights on the perceived quality of the non-horizontal sources was considered comparable in both experiments.

For each source position, the four individually synthesized VAH BRIRs (with V0/β_0, V0/β_{-10}, V0 $\pm$ 15/β_0 and V0 $\pm$ 15/β_{-10}) as well as the (non-individual) HTK and HTS BRIRs were evaluated for four perceptual attributes Source Width, Source Distance, Source Direction, and Overall Quality. The perceptual attribute Reverberance was not considered due to the absence of this attribute for this environment.

5.1 Experiment 2: Results

The consistency test described in Section 3.4 was also used in the present experiment and led to the exclusion of one subject for the perceptual attributes Source Width and Source Distance and two subjects for the perceptual attribute Source Direction. For the perceptual attribute Overall Quality, no subjects were excluded (Fig. 5). It should be noted that three from the four exclusions pertained to one of the subjects with less experience. Figure 6 shows the histograms of differences between the three repetitions. Between 32% and 45% of ratings were identical and between 71% and 84% were within ±1 scale units. A symmetrical distribution of differences with respect to zero difference can be observed for all presentation pairs and attributes.

Figure 9 shows the perceptual evaluations for the four perceptual attributes, three source positions and six BRIR sets. Compared to Experiment 1, the VAH BRIRs with V0/β_0 were rated slightly lower, but still comparable with the HTK and HTS BRIRs, with median values between good and excellent in most cases. Similar to Experiment 1, the VAH BRIRs including non-horizontal directions (V0 $\pm$ 15/β_0, V0 $\pm$ 15/β_{-10}) were rated lower than the VAH BRIRs calculated with only horizontal directions (V0/β_0, V0/β_{-10}) and the HTK and HTS BRIRs. The median values dropped however from between fair and poor in

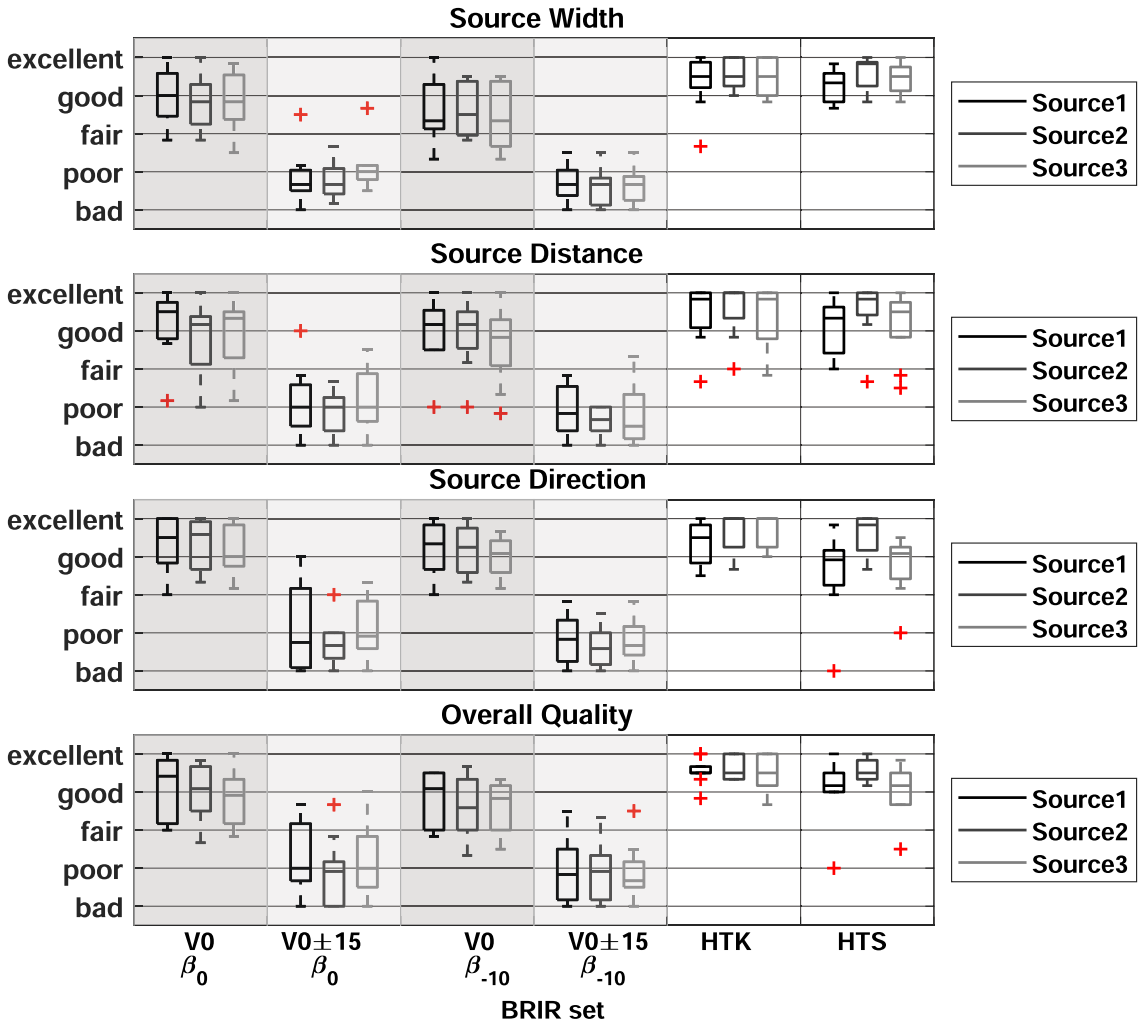

Figure 9. (Experiment 2) Perceptual ratings averaged over three repetitions, for four perceptual attributes, three source positions and six different BRIR sets.

Experiment 1 to around poor and bad. Also, similar to Experiment 1, the VAH BRIRs with V0/β_{-10} were rated slightly lower than the VAH BRIRs with V0/β_0.

The Shapiro-Wilk test of normality revealed that the ratings in Experiment 2 cannot be assumed to be normally distributed for all cases. Therefore, the same non-parametric methods as used in Experiment 1 were applied to the ratings in Experiment 2. A significant effect of the source position was indicated by the Friedman test only for five out of 24 combinations of BRIR sets and perceptual attributes (p-values are shown in Table 2). Therefore, the ratings were again averaged over the three source positions. The averaged ratings are shown in Figure 10. The Friedman test revealed for all attributes a significant effect of the BRIR set ($p < 10^{-4}$). Significantly different BRIR sets (according to the multiple comparisons after Friedman test) are indicated with horizontal lines in Figure 10. Significantly lower ratings were given only to VAH BRIRs with V0 ± 15/β_0 and V0 ± 15/β_{-10}. Similar as in Experiment 1, there were no significant differences between the VAH BRIRs with V0/$\beta_{0,-10}$ and HTK or HTS BRIRs. Also, there were no significant differences between V0/β_0 and V0/β_{-10}.

6 Discussion

6.1 Comparison between auralization and real sound source presentation

For all attributes and for both environments, there were BRIRs for which the median values of the ratings were between good and excellent, i.e. at least 7 and more on the 9-point scale used. As also discussed in [30], if the reference signal is known, subjects tend to avoid the highest point of the scale. Therefore, ratings of between good and excellent were considered as perceptually close to reality. The results suggest that it is possible to have dynamic auralizations with the VAH which are perceived nearly the same as the original acoustical scene, confirming the results that were obtained with simulated BRIRs in [30].

6.2 Low ratings for VAH BRIRs with V0 ± 15 and V0 ± 30

In both experiments, the ratings for VAH BRIRs calculated with horizontal and non-horizontal directions (V0 ± 15/β_0, V0 ± 30/β_0, V0 ± 15/β_{-10} and V0 ± 30/β_{-10})

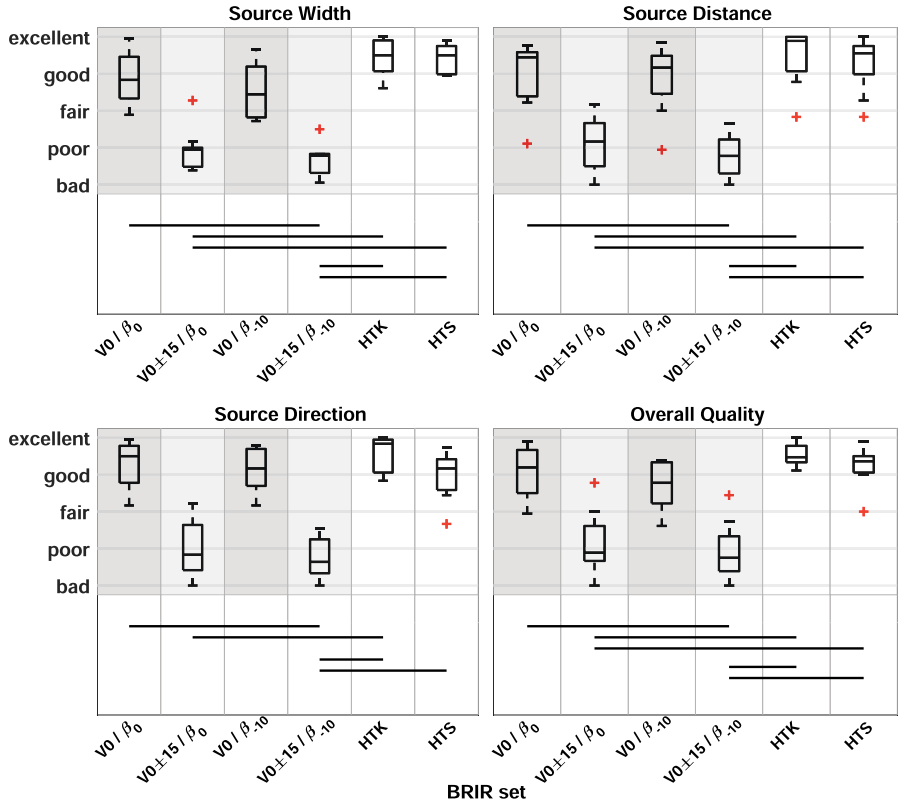

Figure 10. (Experiment 2) Averaged ratings over three source positions for different BRIR sets and perceptual attributes. Significant different ratings are marked with horizontal lines ($p < 0.05$).

were lower than the ratings for VAH BRIRs with V0/β_0 or V0/β_{-10}. This applied to all source positions. For sources in the horizontal plane (e.g., Source 2 in both experiments), one could explain this by the higher resulting SDs at horizontal directions for the case where horizontal and non-horizontal directions were included (compare the left column in Figs. 2c and 2d to Figs. 2a and 2b). However, the lower ratings of VAH BRIRs with V0 ± 15 or V0 ± 30 applied also to sources out of the horizontal plane (Source 4 in Experiment 1 or Source 3 in Experiment 2). These ratings cannot be explained by the SDs at non-horizontal directions, which would predict a better performance of the VAH BRIRs with V0 ± 15 or V0 ± 30. Instead, the ratings seem to be related to the resulting temporal distortion (TD), which is the error in timing of a single frequency component of the BRIR as derived from the phase angle according to:

$$\mathrm{TD}_\zeta(f,\Theta_k) = \frac{\angle \mathbf{w}_\zeta^H(f)\mathbf{d}(f,\Theta_k) - \angle D_\zeta(f,\Theta_k)}{2\pi f}. \quad (7)$$

The resulting TD at two elevations 0° and 15° for syntheses with V0/β_0 and V0 ± 15/β_0 are shown in Figures 11a and 11b. It is important to note that interaural time differences (ITDs) are best perceived below frequencies of 1.5 kHz, as is indicated by the inability to hear ITD changes above this cutoff frequency [43, 44]. Therefore, the high TDs at frequencies below 1.5 kHz are suspected to have led to the lower ratings of the case with horizontal and non-horizontal directions included. It seems that the constrained optimization algorithm sacrificed the phase accuracy to serve the large amount of constraints (216 + 1 in case of V0 ± 15 and V0 ± 30) which were applied to the Spectral Distortion (magnitude error) and mean WNG. Errors in the resulting phase (or TD) will then lead to deviations in the ITDs, which will have impeded the localization ratings. In addition, the ITDs were only implicitly controlled for in the minimization of the cost function while the ILDs were explicitly controlled for as a direct consequence of constraints applied to the Spectral Distortion. As a result, non-matching ITDs and ILDs might have led to a spatial split or a diffuseness of the auditory event [45] or insufficient externalization, which will have impacted the Source Width and Source Distance ratings.

In case of the reverberant environment in Experiment 1, it is also of interest to consider the modified RL_E' (room level (early)), which has been shown to correlate with the perceived apparent source width (ASW) for music [46]. According to this measure, a higher RL_E' corresponds to a larger perceived ASW. Figure 12 shows the RL_E', calculated for the VAH BRIRs of subject 1 and the HTK and HTS using the method described in [46]. The RL_E's in Figure 12 were calculated for the frontal sound source in the lecture room and for horizontal head orientations θ_h between −90° and +90°. The results show higher RL_E' of VAH BRIRs with V0 ± 15/$\beta_{0,-10}$ and V0 ± 30/$\beta_{0,-10}$ compared

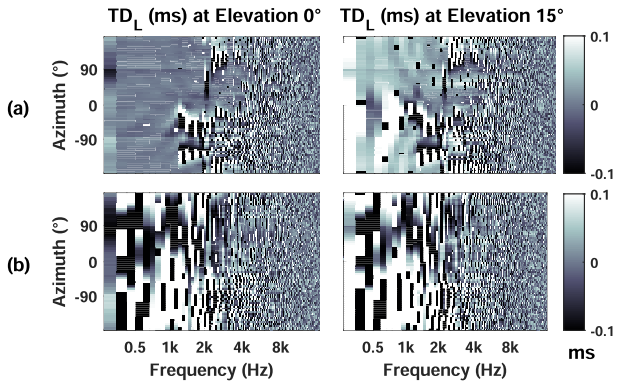

Figure 11. Resulting temporal distortion (TD) at elevations 0° and 15° for spectral weights calculated with (a) V0/β_0, and (b) V0 ± 15/β_0. Results are shown for the left ear of subject 1.

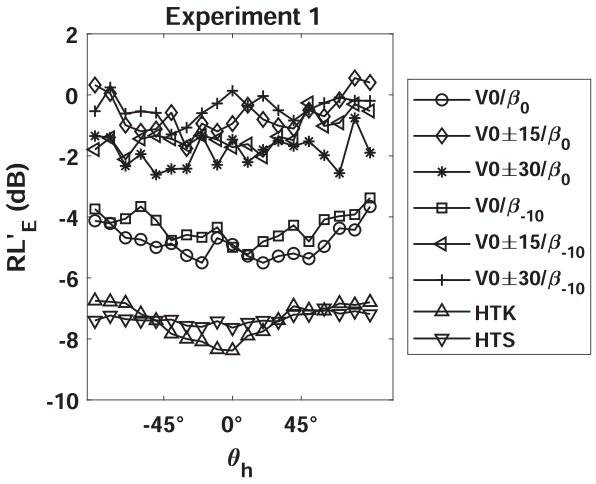

Figure 12. RL'_E calculated for VAH BRIRs of subject 1 and the HTK and HTS BRIRs. RL'_E was calculated for the frontal source in the lecture room.

to VAH BRIRs with V0/$\beta_{0,-10}$ or HTK and HTS BRIRs, which implies that the virtual sources generated with VAH BRIRs with V0 ± 15 or V0 ± 30 were difficult to be perceived at a focused position.

In general, the similarity of the results across perceptual attributes indicated that the synthesis artifacts in VAH BRIRs with V0 ± 15 or V0 ± 30 impacted similarly the quality of the headphone signals with respect to all of the evaluated perceptual attributes.

6.3 The choice of the *P* discrete source directions depending on the application case

In both environments investigated in this study, the VAH BRIRs with V0/β_0 resulted in median ratings between good and excellent for most of the tested source positions and perceptual attributes. Although the resulting SDs at non-horizontal source positions were higher for these VAH BRIRs than with V0 ± 15/$\beta_{0,-10}$ or V0 ± 30/$\beta_{0,-10}$, it seemed that the increasing SDs towards higher frequencies for the BRIRs with V0/β_0 were not very crucial. In addition, the low-frequency TDs were lower with VAH BRIRs with only horizontal directions included. The results imply that it is advantageous to apply the constraints to horizontal directions only.

It must be noted that the advantage of calculating the spectral weights with horizontal source directions is valid for speech signals only, because the SD of VAH BRIRs with V0 at non-horizontal directions only stay within an acceptable range in the frequency range important for speech. In case of applications using signals with a more pronounced high-frequency spectral content, additional audible artifacts are expected to occur at non-horizontal directions.

6.4 The effect of the minimum desired WNG_m

In both experiments, for all source positions and perceptual attributes, the VAH BRIRs with V0/β_{-10} were rated slightly lower than the VAH BRIRs with V0/β_0. Although the effect of microphone self-noise was not evaluated in the same manner using the synthesized or measured BRIRs as it would be using real recordings, a possible mismatch between the measured steering vectors (see Sect. 3.1) and the measured impulse responses (see Sect. 3.2) was present. Since at least four months passed between measuring the steering vectors and measuring the impulse responses in the lecture room and in the anechoic room, it is possible that small deviations in microphone characteristics or positions occurred during this time period. With a lower value of parameter β, the susceptibility of the VAH synthesis to deviations in microphone characteristics increases, which possibly explains the lower ratings given to VAH BRIRs with V0/β_{-10} compared to V0/β_0. The case of $\beta = 0$ dB was perceptually evaluated previously to be a proper choice for the used microphone array in this study [32] and the results in the present study confirmed it. The effect of the parameter β could also be observed for the VAH BRIRs including non-horizontal directions (V0 ± 15/$\beta_{0,-10}$ and V0 ± 30/$\beta_{0,-10}$), although the lower ratings for these VAH BRIRs were dominated by other factors, as discussed in Section 6.2.

6.5 The positive effect of reverberation

The inclusion of reverberation in the binaural synthesis, when congruent with the reverberation of the real room (see Sect. 6.6), can contribute to a better externalization even for the case that non-individual HRTFs of artificial heads are used [30, 47, 48] and help smooth out the deviations to individual HRTFs [8]. The generally higher ratings for VAH BRIRs in Experiment 1 compared to Experiment 2 implied that the synthesis errors of the VAH BRIRs were less audible in the reverberant environment.

The increase of apparent source width in the reverberant environment of Experiment 1 seems to have been particularly in favor of the ratings for Source 3. This source was located at the azimuthal position −112°, which did

not match any of the azimuthal directions considered, regarding the 5° azimuthal resolution of the measured HRTFs and steering vectors. The synthesis at directions other than the ones included in the calculation of spectral weights can be subject to audible artifacts. Although with the relatively close distance of Source 3 to the listener, the direct part of the room impulse response had more energy than the reverberant part, the small ratio of the reverberation included in the measured room impulse response for Source 3 was enough to cover up the potential audible artifacts. Such artifacts would probably have been audible in the anechoic environment, if sources at positions not matching the considered directions had been evaluated in Experiment 2.

The presence of reverberation and reflections were also helpful against the non-individual cues of KEMAR artificial head or the rigid sphere. However, the comparably high ratings given to HTK and HTS BRIRs in the anechoic room suggested that also other factors promoted the high ratings for non-individual BRIRs, which are discussed in Section 6.6.

6.6 The positive effect of head tracking, the compatibility of the auralized and listening rooms, and the presence of visual cues

The similarly high ratings given to HTK and HTS BRIRs and VAH BRIR with V0/β_0 in Experiment 2 are not in accordance with the results of Rasumow et al. [13], where individual binaural presentations generated with the VAH (the same microphone array as used in the present study) in the anechoic room outperformed the presentations generated with a conventional artificial head. The major differences between the study in [13] and the study here were the stimulus and the presentation method. Rasumow et al. evaluated the VAH and artificial head signals with noise bursts in a static scenario, i.e. without head tracking. Broadband test signals appeal a different challenge on the spectral accuracy compared to speech signals. Furthermore, although the advantages of using individual HRTFs are known in a static signal presentation without head tracking (lack of externalization or localization ambiguities [49, 50]), it has been shown that the incorporation of head tracking can significantly reduce the localization ambiguities such as front-back confusions [4] and that the effect of head-tracking is larger than the effect of using individual HRTFs [49, 51]. When using broad band noise signals as in [13] together with head tracking as applied in the present study, the comparison between individually synthesized VAH BRIRs and non-individual BRIRs HTK and HTS is expected to depend on the evaluated perceptual attribute. For example, for non-individual BRIRs HTK and HTS, the perceptual attribute Overall Quality could be rated lower than for individually synthesized VAH BRIRs due to coloration artifacts while other attributes such as Source Direction could still be rated comparable to the VAH BRIRs due to the incorporation of head tracking. With speech signals used in the present study however, the dynamic presentation of the signals was advantageous for the perceived quality of non-individual binaural signals with respect to all of the evaluated perceptual attributes.

In addition, other features promoted the quality of the signals generated with VAH, HTK and HTS BRIRs in this study. For both experiments, the listening test was performed in the same environment that was also auralized, with all perceptual cues preserved as they were during the impulse response measurements. A discrepancy between the auralized room and the listening room can impact the externalization or the perceived distance of the sound source negatively [52, 53]. Another relevant feature was the visual information about the sources and their positions in the room. The knowledge of the source position can help suppress front-back confusions and improve the externalization. In addition, the presence of visual information can draw the acoustically perceived source position to the visual one [54].

At any time in everyday life, the surrounding environment is being perceived and evaluated based on the information available from different modalities in accordance with each other. The present study also offered a high consistency between the acoustical and visual features. Regarding the high perceptual ratings given to non-individual BRIRs of HTK or HTS, one can question the need for individualizing the binaural recordings, if the head-tracked binaural presentation, applied to less critical signals such as speech, can maintain such a consistency, especially for cases where no external reference is provided.

6.7 VAH vs. traditional artificial head

As discussed in Section 6.6, the possibility of applying head-tracking to the non-individual binaural signals of the conventional artificial head can be expected to improve the perceptual quality. Regarding the fact that the conventional artificial heads do not normally offer the possibility of a dynamic head-tracked presentation in their standard applications, the incorporation of head tracking constitutes the great advantage of the VAH technology against these conventional artificial heads. Although the spectral weights for a high number of head orientations requires a high number of calculations, these spectral weights are calculated only once and can then be applied to any recording. The comparable perceptual ratings given to VAH BRIRs with V0/β_0 and HTK or HTS together with the provided ability of the VAH to allow dynamic auralizations confirmed that the VAH is the more promising alternative for head-tracked auralizations of different environments with a realistic signal such as speech.

7 Conclusion

In this study, the virtual artificial head (VAH) was used to synthesize individual binaural room impulse responses (BRIRs) in two acoustically different environments (lecture room and anechoic room). VAH spectral weights were calculated for 185 head orientations (37 horizontal × 5 vertical), individually for each listener, using different sets of

parameters. Individual BRIRs were synthesized by filtering the room impulse responses measured with the VAH with the FIR filters corresponding to the inverse Fourier transform of the spectral weights.

The results of the perceptual evaluations suggest that realistically (i.e. perceptually close to the original scenario) sounding head-tracked auralizations of speech can be realized using the VAH technology. This was shown for two different acoustical environments and for sources in and out of the horizontal plane. The choice of the discrete source directions included in the calculation of the spectral weights is critical for the quality of the synthesis. According to the perceptual results, it was advantageous to include directions from the horizontal plane only. A total of 72 horizontal directions together with 5° resolution for the horizontal head orientations was sufficient to achieve good perceptual results with the VAH. The slightly higher perceptual results for the reverberant environment indicated the positive effect of reverberation in masking the synthesis errors and thus improving the perceptual quality of the synthesis with the VAH.

The results also showed that the resulting mean White Noise Gain (WNG_m), as a measure for robustness, can as well impact the quality of the binaural signals generated with the VAH. In general, it is advisable to avoid low resulting WNG_m in order to increase the robustness of the microphone array against e.g. changes in microphone positions or microphone self-noise.

Non-individual BRIRs measured with a conventional artificial head or a simple rigid sphere can also result in highly realistic auralizations of speech, provided that head tracking with sufficiently many head orientations is employed. This means that different head orientations have to be accounted for by repeating the BRIR measurements. This will only rarely be an option in BRIR measurements and not be possible in live recordings. It is still interesting to note that individual BRIRs are not necessarily required for the case that the binaural speech signals can be presented dynamically.

The success of the VAH by including only horizontal source directions, as reported in this study, applies to the tested speech signal or signals with comparable spectral content only. When listening to broadband signals, the inclusion of non-horizontal source directions is expected to be more critical for preserving the synthesis accuracy at positions outside the horizontal plane. In addition, when using other test signals, the appropriateness of the spatial resolution for different head orientations in this study (5°) should be verified as well. More accurate statements with this regard require further perceptual evaluations.

It would also be interesting to investigate the extent to which the effect of the head tracking and visual cues contributed to the results, by performing perceptual experiments in the absence of these features.

Conflict of interest

Author declared no conflict of interests.

Acknowledgments

This work was funded by Bundesministerium für Bildung und Forschung under grant No. 03FH021IX5. We would like to thank our subjects for their participation in the study.

References

1. M. Vorländer: Auralization. Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality, 1st ed. Springer, Berlin, Heidelberg, 2008.
2. L. Savioja, U.P. Svensson: Overview of geometrical room acoustic modeling techniques. Journal of the Acoustical Society of America 138 (2015) 708–730.
3. T. Wendt, S. van de Par, S.D. Ewert: A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation. Journal of the Audio Engineering Society 62 (2014) 748–766.
4. D.R. Begault, E.M. Wenzel, M.R. Anderson: Direct comparison of the impact of head tracking, reverberation and individualized head-related transfer functions on the spatial perception of a virtual speech source. Journal of the Audio Engineering Society 49 (2001) 904–916.
5. 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.
6. E. Hendrickx, P. Stitt, J.C. Messonnier, J.M. Lyzwa, B.F.G. Katz, C. deBoishéraud: Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis. Journal of the Acoustical Society of America 141 (2017) 2011–2023.
7. A. Lindau, S. Weinzierl: Assessing the plausibility of virtual acoustic environments. Acta Acustica United with Acustica 98 (2012) 804–810.
8. F. Brinkmann, A. Lindau, S. Weinzierl: On the authenticity of individual dynamic binaural synthesis. Journal of the Acoustical Society of America 142 (2017) 1784–1795.
9. S. Nagel, P. Jax: Dynamic Binaural Cue Adaptation. International Workshop on Acoustic Signal Enhancement, Tokyo, Japan, 2018.
10. C. Pörschmann, P. Stade, J.M. Arend: Binaural auralization of proposed room modifications based on measured omnidirectional room impulse responses. Proceedings of Meetings on Acoustics 30 (2017) 015012.
11. V.R. Algazi, R.O. Duda, D.M. Thompson: Motion-tracked binaural sound. Journal of the Audio Engineering Society 52 (2004) 1142–1156.
12. A. Lindau, S. Roos: Perceptual evaluation of discretization and interpolation for motion-tracked binaural (MTB) recordings, in Proc. of the 26th Tonmeistertagung, VDT International Convention, Leipzig, Germany. 2010, pp. 680–701.
13. 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. Journal of the Audio Engineering Society 65 (2017) 448–459.
14. S. Sakamoto, S. Hongo, T. Okamoto, Y. Iwaya, Y. Suzuki: Sound-space recording and binaural presentation system based on a 252-channel microphone array. Acoustical Science and Technology 36 (2015) 516–526.
15. S. Delikaris-Manias, J. Vilkamo, V. Pulkki: Parametric binaural rendering utilizing compact microphone arrays, in IEEE ICASSP, 19–24 April 2015, South Brisbane, QLD, Australia. 2015, pp. 629–633.

16. J. Chen, B.D. Van Veen, K.E. Hecox: External ear transfer function modeling: A beamforming approach. Journal of the Acoustical Society of America 92 (1992) 1933–1944.
17. B. Rafaely: Analysis and design of spherical microphone arrays. IEEE Transactions on Speech and Audio Processing 13 (2005) 135–143.
18. B. Bernschütz: Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording. Ph.D. thesis, Technische Universität Berlin, 2016.
19. J. Ahrens, C. Andersson: Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre. Journal of the Acoustical Society of America 145 (2019) 2783–2794.
20. J. Atkins: Robust beamforming and steering of arbitrary beam patterns using spherical arrays, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 16–19, New Paltz, NY. 2011, pp. 237–240.
21. Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, B. Rafaely: Spectral equalization in binaural signals represented by order-truncated spherical harmonics. Journal of the Acoustical Society of America 141 (2017) 4087–4096.
22. C. Hold, H. Gamper, V. Pulkki, N. Raghuvanshi, I.J. Tashev : Improving binaural ambisonics decoding by spherical harmonics domain tapering and coloration compensation, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. 2019, pp. 261–265.
23. M. Zaunschirm, M. Frank, F. Zotter: Binaural rendering with measured room responses: First-order ambisonic microphone vs dummy head. Applied Science 10 (2020) 1631.
24. J. Ahrens: Perceptual Evaluation of Binaural Auralization of Data obtained from the Spatial Decomposition Method, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 20–23, New Paltz, NY. 2019, pp. 65–69.
25. V. Pulkki: Spatial sound reproduction with directional audio coding. Journal of the Audio Engineering Society 55 (2007) 503–516.
26. A. Politis, L. McCormack, V. Pulkki: Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing, in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 15–18, New Paltz, NY. 2017.
27. E. Rasumow, M. Hansen, S. van de 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.
28. M. Fallahi, M. Hansen, S. Doclo, S. van de Par, V. Mellert, D. Püschel, M. Blau: High spatial resolution binaural sound reproduction using a virtual artificial head, in Proc. Fortschritte der Akustik – DAGA 2017, Kiel. 2017, pp. 1061–1064.
29. J. Blauert: Spatial Hearing. The psychophysics of human sound localization. Revised Ed., MIT Press, Cambridge, MA, Chapter 2, 1997, pp. 36–200.
30. M. Blau, A. Budnik, M. Fallahi, H. Steffens, S.D. Ewert, S. van de Par: Toward realistic binaural auralizations – perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario, Acta Acustica United with Acustica 5 (2021) 8.
31. 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. Journal of the Acoustical Society of America 135 (2014) 2012–2025.
32. E. Rasumow, M. Blau, M. Hansen, S. Doclo, S. van de Par, V. Mellert, D. Püschel: The impact of the white noise gain (WNG) of a virtual artificial head on the appraisal of binaural sound reproduction, in Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany. 2014.
33. A. Lindau, S. Weinzierl: On the spatial resolution of virtual acoustic environments for head movements in horizontal, vertical and lateral direction, in Proc. of the EAA Symposium on Auralization, Espoo, Finland. 2009.
34. P. Majdak, P. Balazs, B. Laback: Multiple exponential sweep method for fast measurement of head-related transfer functions. Journal of the Audio Engineering Society 55 (2007) 623–637.
35. A. Novák, L. Simon, F. Kadlec, P. Lotton: Nonlinear system identification using exponential sweep-sine signal. IEEE Transactions on Instrumentation and Measurement 59 (2010) 2220–2229.
36. J. Poppitz, M. Blau, M. Hansen: Entwicklung und Evaluation eines Systems zur Messung individueller HRTFs in privater Wohn-Umgebung, in Proc. Fortschritte der Akustik - DAGA 2016, Aachen. 2016, pp. 812–815.
37. O. Kirkeby, P.A. Nelson: Digital filter design for inversion problems in sound reproduction. Journal of the Audio Engineering Society 47 (1999) 583–595.
38. H. Jaeger, J. Bitzer, U. Simmer, M. Blau: Echtzeitfähiges binaurales Rendering mit Bewegungssensoren von 3-D Brillen, in Proc. Fortschritte der Akustik – DAGA 2017, Kiel. 2017, pp. 1130–1133.
39. ITU P.800: ITU-T Recommendations. https://www.itu.int/ITU-T/recommendations/index.aspx (Last viewed March 15, 2021).
40. International Phonetic Association and International Phonetic Association Staff: Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press, 1999.
41. L.J. Cronbach: Coefficient Alpha and the Internal Structure of Tests. Psychometrika 16 (1951) 297–334.
42. S. Siegel, N.J. Castellan: Non Parametric Statistics for Behavioural Sciences. Second ed., McGraw-Hill, Inc. 1988, Chapter 7, pp. 168–189.
43. R.Y. Litovsky, M.J. Goupell, R.R. Fay, A.N. Popper: Binaural Hearing. With 93 Illustrations. 1st ed., Springer, Cham, 2021.
44. B. Grothe, M. Pecka, D. McAlpine: Mechanisms of sound localization in mammals. Physiological Reviews 90 (2010) 983–1012.
45. W. Gaik: Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. Journal of the Acoustical Society of America 94 (1993) 98–110.
46. M. Blau: Correlation of apparent source width with objective measures in synthetic sound fields. Acta Acustica United with Acustica 90 (2004) 720–730.
47. F. Wendt, R. Höldrich, M. Marschall: How binaural room impulse responses influence the externalization of speech, in Proc. Fortschritte der Akustik – DAGA 2019, Rostock. 2019, pp. 627–630.
48. F. Völk, F. Heinemann, H. Fastl: Externalization in binaural synthesis: effects of recording environment and measurement procedure, in Acoustics 08, Paris. 2008, pp. 6419–6424.
49. 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).
50. H. Møller, M.F. Sørensen, C.B. Jensen, D. Hammershøi: Binaural technique: Do we need individual recordings? Journal of the Audio Engineering Society 44 (1996) 451–469.

51. D. Ackermann, F. Fiedler, F. Brinkmann, M. Schneider, S. Weinzierl: On the acoustic qualities of dynamic pseudobinaural recordings. Journal of the Audio Engineering Society 68 (2020) 418–427.
52. S. Werner, F. Klein, T. Mayenfels, K. Brandenburg: A summary on acoustic room divergence and its effect on externalization of auditory events, in Proc. IEEE 8th International Conference on Quality of Multimedia Experience (QoMEX). 2016.
53. J.C. Gil-Carvajal, J. Cubick, S. Santurette, T. Dau: Spatial hearing with incongruent visual or auditory room cues. Scientific Reports 6 (2016) 37342.
54. C.V. Jackson: Visual factors in auditory localization. Quarterly Journal of Experimental Psychology 5 (1953) 52–65.

Cite this article as: Fallahi M. Hansen M. Doclo S. van de Par S. Püschel D, et al. 2021. Evaluation of head-tracked binaural auralizations of speech signals generated with a virtual artificial head in anechoic and classroom environments. Acta Acustica, **5**, 30.