PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Individualized dynamic binaural auralization of classroom acoustics using a virtual artificial head

Mina FALLAHI[1]; Martin HANSEN[1]; Simon DOCLO[2,4]; Steven VAN DE PAR[2,4]; Dirk PUESCHEL[3]; Matthias BLAU[1,4]

[1]Institut für Hörtechnik und Audiologie, Jade Hochschule, Oldenburg, Germany
[2]Dept. Medical Physics and Acoustics, University of Oldenburg, Germany
[3]Akustik Technologie, Göttingen, Germany
[4]Cluster of Excellence Hearing4All, Germany

Abstract

Binaural Room Impulse Responses (BRIRs) are typically used for binaural auralizations, as they carry both room acoustic information as well as acoustical properties of the individual listener. In a dynamic reproduction scenario, i.e. with head movements of the listener taken into account, BRIRs need to be measured for a large number of listener head orientations, which can be very time-consuming. In this study, a planar microphone array with 24 microphones, referred to as Virtual Artificial Head (VAH) was used to measure Room Impulse Responses (RIRs) for a single listener position for different source positions in a lecture room. VAH filter coefficients for different head orientations were calculated and applied to the measured RIRs. The VAH filter coefficients were calculated by solving a constrained optimization problem, i.e. by minimizing a narrow-band least-squares cost function subject to constraints on spectral magnitude error and the mean White Noise Gain. The quality of the synthesized individual BRIRs was evaluated in a subjective listening test based on a dynamic binaural scenario. The results showed that a typical reverberant room can be dynamically auralized with the VAH for speech signals with perceptually convincing agreement to real loudspeaker presentation in the room.

Keywords: Binaural Room Impulse Response (BRIR), Virtual Artificial Head (VAH), dynamic binaural synthesis

1 INTRODUCTION

The signals arriving at the two ears include important cues on sound reflection and diffraction caused by the listener's head, torso and external ear. These cues, which can be captured by Head-Related Transfer Functions (HRTFs), are crucial for a spatial perception of the surrounding acoustic world. The binaural technology commonly makes use of the so-called artificial heads, copies of the average human head and torso with microphones in the ear canals, to capture the spatial properties of the sound in a recorded signal. HRTFs or their equivalence in the time domain referred to as Head Related Impulse Responses (HRIRs) are known to be individual and should preferably be determined individually. Therefore, the non-individual anthropometric geometries of these artificial heads can lead to perceptible deficiencies. In another application of the binaural technology, a given room can be auralized binaurally by convolving an anechoic signal with Binaural Room Impulse Responses (BRIRs). BRIRs combine the information contained in HRIRs with the acoustical information of the room (e.g. lateral reflections and reverberation). With binaural auralization, naturally sounding representations of different rooms can be created, which establish an appropriate tool for the investigation and comparison of different acoustical environments. Furthermore, including head movements during binaural signal playback with headphones improves the localization accuracy [1], helps to promote externalization [2] and plays an important role in preserving the plausibility [3] and authenticity [4] of the virtual acoustic environment. In a dynamic binaural synthesis, the listener's head movements are tracked in real-time to choose the BRIR corresponding to the current head orientation to be convolved with the signal. This means, however, that in order to auralize different reverberant environments dynamically, the BRIRs should be measured for different head orientations in each environment, which is very time consuming, especially if the BRIRs are supposed to be measured individually.

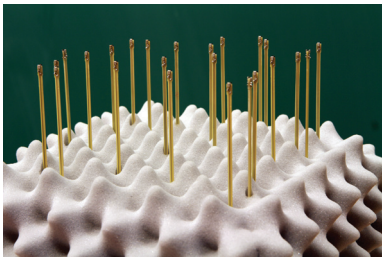

Figure 1. Virtual Artificial Head (VAH) used in this study: planar microphone array with 24 microphones [5]

As an alternative, the Virtual Artificial Head (VAH) can be employed which has been shown to capture spatial audio signals for headphone representations in a perceptually convincing manner [5]. The VAH consists of a microphone array with N spatially distributed microphones with digital filters for each microphone, forming a filter-and-sum beamformer. The same recording made with the VAH can be individualized post-hoc by mimicking the directivity pattern of the individual HRTFs. This is done by applying the individually calculated filter coefficients as complex-valued weighting factors to the signals of the N microphones. The major advantage of the VAH is that the filter coefficients can be calculated for different head orientations. This allows for head tracking during signal playback without the need for changing the VAH's orientation during the recording. The VAH technology for a measured and simulated room was evaluated in listening tests and found to be perceptually convincing for dynamic binaural auralizations in [6], using simulated steering vectors. In the present study, real measurements were accomplished with the VAH in the same room as in [6] to assess the extent to which BRIRs synthesized with the VAH can be used for dynamic binaural room auralizations. The results were perceptually evaluated in a listening test with comparisons to real loudspeaker presentations in a typical classroom.

2 FILTER COEFFICIENTS FOR THE VIRTUAL ARTIFICIAL HEAD (VAH)

2.1 Calculation of VAH filter coefficients in a constrained optimization problem

The directivity pattern $H(f,\theta)$ of VAH at frequency f and direction θ is defined as

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

with the $N \times 1$ steering vector $\mathbf{d}$, defined as the free-field acoustical transfer function between the source at any given direction θ and the N microphones. The $N \times 1$ complex-valued vector $\mathbf{w}(f)$ of Filter Coefficients (FCs) consists of the weighting factors for the N microphone signals. In order to synthesize the desired directivity pattern $D(f,\theta_k)$ of individual left and right HRTFs at a total of P discrete directions θ_k, $k = 1,2,...,P$, FCs can be calculated by minimizing a narrow-band least-squares cost function J_{LS}, defined as

$$J_{LS}(\mathbf{w}(f)) = \sum_{k=1}^{P} |H(f,\theta_k) - D(f,\theta_k)|^2. \tag{2}$$

In order to achieve small synthesis spectral deviations at all P directions, the minimization of J_{LS} was performed subject to constraints imposed to the resulting Spectral Distortion (SD) at each of the directions θ_k, by setting an upper and a lower limit, L_{Up} and L_{Low}, i.e. for all k

$$L_{\mathrm{Low}} \leq SD(f,\theta_k) = 10\lg \frac{|\mathbf{w}^H(f)\mathbf{d}(f,\theta_k)|^2}{|D(f,\theta_k)|^2}\mathrm{dB} \leq L_{\mathrm{Up}}. \tag{3}$$

In addition, in order to guarantee the robustness of VAH against microphone self-noise or deviations in microphone characteristics and positions, a minimum desired value β was set for the resulting *mean* White Noise

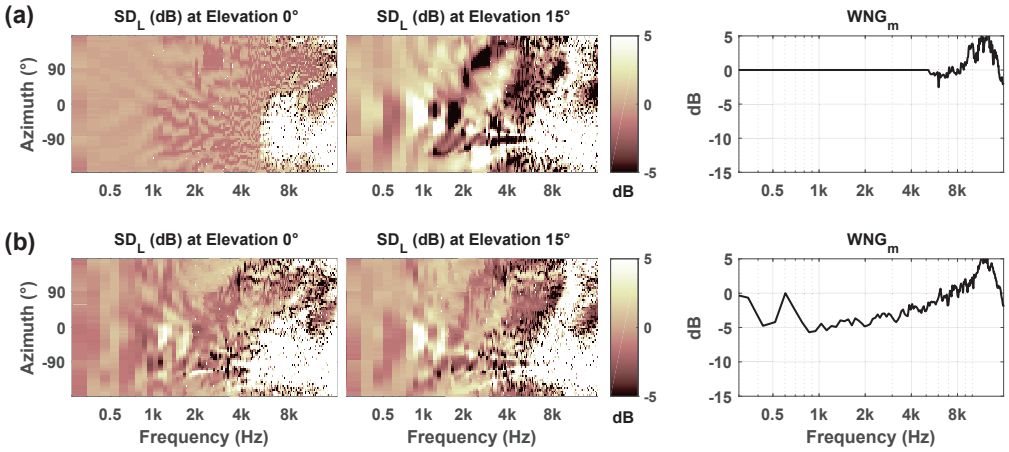

Figure 2. Resulting Spectral Distortion (*SD*) for synthesizing left HRTFs at 0° and 15° elevations as well as resulting left $\mathrm{WNG_m}$, using FCs calculated with: (a) $P = 72$ horizontal directions and $\beta = 0$ dB, (b) $P = 3 \times 72 = 216$ directions from elevations $-15°$, 0° and 15° and $\beta = 0$ dB.

Gain, WNG_m, which is defined as the ratio between mean output power of the microphone array over all P directions and the output power of the spatially uncorrelated white noise [7], i.e.

$$WNG_m = 10\lg\left(\frac{1}{P}\sum_{k=1}^{P}\frac{|\mathbf{w}^H(f)\mathbf{d}(f,\theta_k)|^2}{\mathbf{w}^H(f)\mathbf{w}(f)}\right)\mathrm{dB} \geq \beta. \tag{4}$$

For solving the constrained optimization problem of minimizing J_{LS} subject to constraints defined in Eqs. 3 and 4, an iterative Interior-Point optimization algorithm was used. The solutions proposed in [7] were used as the initial values.

2.2 Constraints and parameters for the calculation of the VAH filter coefficients

According to Eqs. 2-4, the performance of the VAH with respect to meeting the constraints depends on a variety of parameters such as constraint parameters L_{Up}, L_{Low} and β, the microphone array topology and the number P of discrete directions included in the calculation of the FCs. Applying different constraint parameters to two simulated microphone arrays of different topologies, it was shown in [8] that with properly chosen constraint parameters and array topology, the VAH can perceptually outperform a traditional artificial head for sources in the horizontal plane for musical content with respect to overall audio quality. However, since the microphone arrays were simulated and considered as perfectly robust and noise-free, the effect of different values of resulting $\mathrm{WNG_m}$ could not be assessed in [8]. In addition, the results were evaluated only for free-field conditions. In contrast, in the current study the planar microphone array with 24 microphones shown in Figure 1 was used for measurements in a reverberant room. The two constraint parameters L_{Low} and L_{Up} were chosen as -1.5 dB and 0.5 dB, respectively, leading to a maximum deviation of 2 dB in the resulting Interaural Level Differences at all P directions, as was done earlier in [8].

The values chosen for the two other parameters, i.e. β and P, impacted the extent to which the resulting *SD* would remain between the two desired limits of -1.5 dB and 0.5 dB. For example, Figure 2a shows the resulting *SD* and $\mathrm{WNG_m}$ for synthesizing left HRTFs with FCs calculated with $\beta = 0$ dB and $P = 72$ directions in the horizontal plane (5° azimuthal resolution). A minimum value of 0 dB for the resulting $\mathrm{WNG_m}$ and -1.5 dB $\leq SD \leq$ 0.5 dB at P=72 directions in the horizontal plane could be satisfied well up to 6 kHz. However, for non-horizontal directions such as at 15° elevation the resulting *SD* increased clearly. On the other hand,

including P = 216 directions (3×72) from elevations $-15°$, $0°$ and $15°$ improved the resulting *SD* at elevation $15°$, but deteriorated at the same time the resulting $\mathrm{WNG_m}$ and the *SD* at horizontal directions, as shown in Figure 2b. If more elevations were included in the calculation of the VAH FCs, the resulting *SD* and $\mathrm{WNG_m}$ would deteriorate even more for horizontal directions. Thus, the question arises, whether it is preferable to target at high accuracy and robustness in the horizontal plane while ignoring non-horizontal directions or to target moderate accuracy and robustness which are then more evenly distributed over all elevations. For this study, three cases for P in combination with two cases for β, as listed in Table 1, were considered, resulting in a total of six sets of FCs.

In order to enable a dynamic binaural presentation, each of the six sets of FCs were calculated for 37×5 = 185 head orientations (azimuth angles $-90°$ to $+90°$ in $5°$ steps and elevations $-15°$ to $+15°$ in $7.5°$ steps). For a given head orientation θ_h, $h \in 1,2,...,P$, FCs were calculated by taking the $D(f,\theta_k)$, $k = 1,2,...,P$, and the shifted steering vectors $\mathbf{d}(f,\theta_{k'})$ with $k' = h,h+1,...,P,1,2,...h-1$ into Eqs.(1) to (4).

Table 1. Overview of values chosen for the parameters P and β.

P = 1 × 72	(Elevation: $0°$) labeled as **El0**	β = 0 dB labeled as $\boldsymbol{\beta_0}$
P = 3 × 72 = 216	(Elevations: $-15°$, $0°$, $15°$) labeled as **El0±15**	
P = 3 × 72 = 216	(Elevations: $-30°$, $0°$, $30°$) labeled as **El0±30**	β = -10 dB labeled as $\boldsymbol{\beta_{-10}}$

3 METHODS

3.1 Measurement room

The room chosen for this study, which was the same as in [6], was a lecture room (7.12m ×11.94m × 2.98m) with an average reverberation time of 0.58 s. A listener position was selected with ears at 1.30m height from the floor, as shown in Figure 3a. The figure also shows the four source positions that were considered in the room: Source 1 (Genelec type 8030c) was located ahead of the listener slightly higher than the ears, representing a lecturer standing in front. Source 2 and Source 3 (Genelec type 8030b) were located left and behind the listener at the right side, respectively, both at the same height as the ears. Source 4 (Event active studio monitor 20/20 bas V3) was mounted in front of the room on the right and at an elevation of about $20°$.

3.2 BRIR acquisition

In the next step, the VAH was placed at the listener position and the Room Impulse Responses (RIRs) for all four source positions were captured for each of the 24 microphones of the VAH (Figure 3 b). These RIRs were then filtered with the previously mentioned six sets of FCs, resulting in six versions of individually synthesized BRIRs, referred to as VAH BRIRs. BRIRs were also measured with the KEMAR artificial head as well as with a head-sized rigid sphere (radius = 8.5 cm) with two microphones positioned at $\pm 100°$ on the equator, referred to as KEMAR BRIRs and Sphere BRIRs, respectively (see Figure 3 c-d). In order to enable dynamic binaural representations, the KEMAR- and Sphere BRIRs had to be measured 37 times for 37 orientations of the artificial head and the rigid sphere (azimuthal head orientations -90° to 90°, in 5° steps). Note that KEMAR's head was rotated relative to the shoulders which kept a constant orientation. All RIRs were measured with a sampling rate of fs = 44100 Hz, using the Multiple Exponential Sweep Method (MESM) [9] with sweeps of 20 s duration, ranging from 20 Hz to fs/2 with 4 s shift between subsequent excitations.

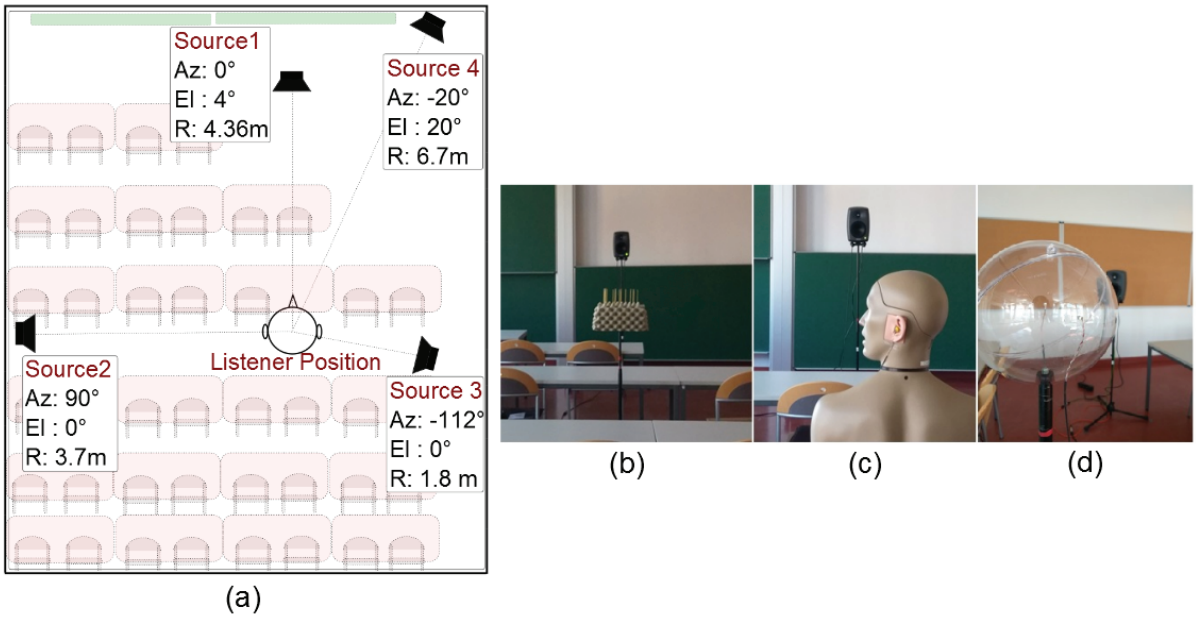

Figure 3. (a) Listener position and the four source positions in the lecture room (Az: azimuth, El: elevation, R: distance to listener. (b) VAH, (c) Kemar artificial head, and (d) rigid sphere at listener position in the room.

3.3 LISTENING TEST

To assess the perceptual quality of the six VAH BRIRs as well as the KEMAR- and Sphere BRIRs, a listening test with dynamic binaural synthesis was performed. A total of 9 normal-hearing subjects (three female, six male) took part in the test. For all of them, individually measured HRTFs and Head Phone Transfer Functions (HPTFs) as well as six sets of individually calculated FCs were available. The headphone (Sennheiser HD800) was equipped with a head tracker attached to the top bow as well as a push button to switch back and forth from headphone to loudspeaker presentation, as implemented in [6]. The listening test took place in the same room as the measurement room. Subjects sat at the listener position during the test and were asked to rate different headphone presentations, generated either with VAH BRIRs or KEMAR- and Sphere BRIRs, in comparison to the presentation via loudspeakers in the room (the reference condition). The subjects were not informed about the specific BRIR synthesis condition that was presented. The stimulus was a dry recorded speech utterance of 15 s duration, spoken by a female speaker, which was convolved with the different BRIRs as well as the inverse HPTFs prior to presentation via headphones. The evaluation was performed with respect to five separate attributes: Reverberance, Source Width, Source Distance, Source Direction and Overall Quality. Subjects gave their ratings on a 9-point scale covering five German labels schlecht (bad), dürftig (poor), ordentlich (fair), gut (good) and ausgezeichnet (excellent) with four equidistant intermediate scale points. Each source direction appeared three times in the test in a randomized order and subjects were allowed and encouraged to switch freely between all of the different headphone signals and/or between loudspeaker and headphone presentations.

4 RESULTS AND DISCUSSION

The perceptual ratings of all subjects for different perceptual attributes and source positions regarding different BRIRs are shown in Figure 4. Almost for all perceptual attributes and source positions the VAH BRIRs with El0/β_0 and the KEMAR and Sphere BRIRs were rated similarly high, with median ratings between good and

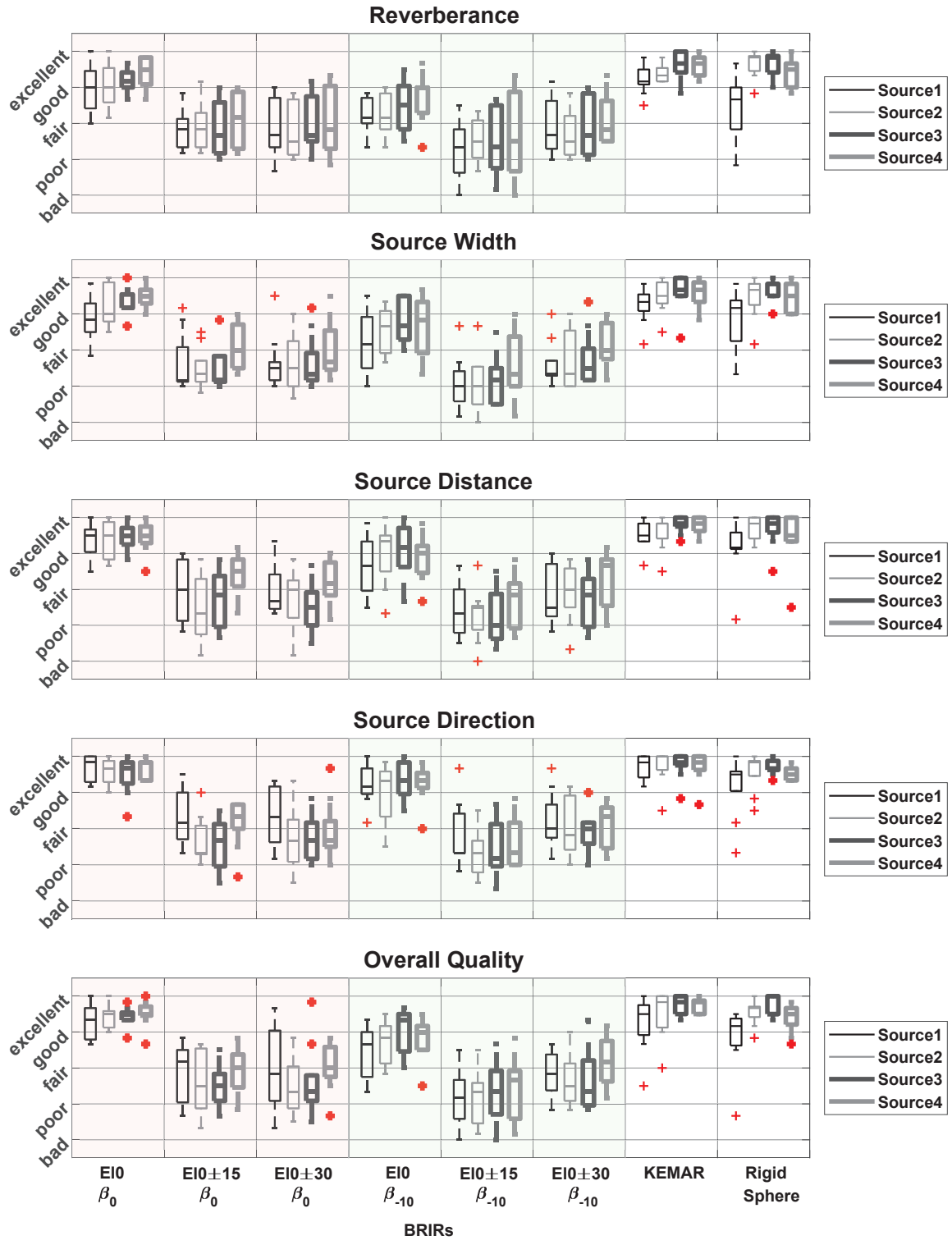

Figure 4. Results of perceptual evaluations of 9 subjects (averaged over three repetitions) with respect to five attributes and four source positions, for 8 different BRIR versions (horizontal axis).

excellent. The two exceptions were the ratings given to Sphere BRIRs with respect to Reverberance, and the ratings given to VAH BRIRs with El0/β_0 with respect to Source Width, both for the signal corresponding to Source 1. In comparison, generally lower ratings were given everywhere to VAH BRIRs with more directions included in the calculation of FCs, i.e. El0$\pm$15 and El0$\pm$30, regardless of β. This is in accordance with the poorer synthesis accuracy in the horizontal plane, as already discussed as an example in Figure 2, and indicates that accuracy and robustness for directions in the horizontal plane are more important than for non-horizontal directions in a classroom scenario, eventhough not all sources were in the horizontal plane. VAH BRIRs with El0/β_{-10} were in general rated lower than VAH BRIRs with El0/β_0. Since for the calculation of FCs for El0/β_{-10} the constraint imposed to the resulting $\mathrm{WNG_m}$ was relaxed to allowable values down to -10 dB, the sensitivity of the VAH to deviations in microphone characteristics increased for VAH BRIRs with β_{-10}. Regarding the time lapse of about four months between the measurement of the steering vectors for the calculation of FCs and the recordings performed in the room, it seems likely that small unavoidable deviations in microphone positions and/or sensitivities during this time period might have led to audible synthesis artifacts in the VAH BRIRs with El0/β_{-10}. As a result, it is advisable to choose higher values for β.
Despite the different geometries of the Kemar artificial head and the rigid sphere compared to anthropometric geometries of individual subjects, high ratings were given almost in every case to the KEMAR- and Sphere BRIRs. We suspect that this is due to the presence of reflections in the reverberant room which improved the externalization and helped mask the deficiencies related to non-individual BRIRs (see for example [4]). Another important aspect was that the listening test took place in the original room and hence, all other cues including visual ones were perfectly available. This made it for example less likely that visual cues that are inconsistent with auditory cues lead to in-head-localization or front-back confusions. Nevertheless, the ratings given to VAH BRIRs with El0/β_0 were similarly as good as ratings on KEMAR- and Sphere BRIRs. Applying the Friedman test with post-hoc multiple comparisons revealed in some cases significantly better ($p<0.05$) ratings given to KEMAR- and Sphere BRIRs compared to some of the other VAH BRIRs, but there were no significant differences between ratings given to VAH BRIRs with El0/β_0 and KEMAR- and Sphere BRIRs for any of the perceptual attributes or source positions. Regarding the impractical effort taken to measure the KEMAR- and Sphere BRIRs for many different head orientations, and considering the comparable perceptual results given to VAH BRIRs with El0/β_0, the VAH technology seems to offer the more promising alternative for dynamic binaural auralizations of commonly used acoustical environments (lecture room) with realistic signals (speech).

5 SUMMARY AND CONCLUSION

In this study, individual binaural room impulse responses (BRIRs) in a lecture room were synthesized for 185 head orientations with the virtual artificial head (VAH). The synthesized BRIRs of different head orientations were used to auralize the room dynamically. The quality of the synthesized BRIRs was evaluated regarding different perceptual attributes in a listening test in comparison to the real loudspeaker playback in the room. The results showed that it is possible to auralize a typical reverberant room dynamically with the VAH for speech signals, with perceptually high agreement to the real loudspeaker signal. The accuracy and robustness for directions in the horizontal plane is very important for sources in and near the horizontal plane. The results showed also that the need for BRIR individualization might not be crucial in a reverberant room. It should be noted, that this outcome relates only to the case of dynamic binaural representation. To which extent the effect of head tracking had an influence on the evaluations should be addressed in a separate experiment. Nevertheless, the major advantage of the VAH with respect to the possibility of applying head tracking could be demonstrated in this study. Further investigations should concern the performance of the VAH in other acoustical environments, for other source positions and using other signals such as music or broadband noise, alongside the enhancement of the VAH with respect to the microphone array topology.

ACKNOWLEDGEMENTS

This work was funded by Bundesministerium für Bildung und Forschung under grant no. 03FH021IX5.

REFERENCES

[1] Begault, D. R., Wenzel, E. M., Anderson, M.R. Direct comparison of the impact of head tracking, reverberation and individualized head-related transfer functions on the spatial perception of a virtual speech source. *J. Audio Eng. Soc.*, 49(10), pp. 904-916, 2001.

[2] Brimijoin, W. O., Boyd, A. W., Akeroyd, M. A. The contribution of head movement to the externalization and internalization of sounds. *PLOS ONE*, 8(12), 2013.

[3] Lindau, A., Weinzierl, S. Assessing the plausibility of virtual acoustic environments. *Acta Acustica united with Acustica*, 98, pp. 804-810, 2012.

[4] Brinkmann, F., Lindau, A., Weinzierl, S. On the authenticity of individual dynamic binaural synthesis. *J. Acoust. Soc. Am.*, 142(4), pp. 1784-1795 2017.

[5] Rasumow, E., Blau, M., Doclo, S., van de Par, S., Hansen, M., Püschel, D., Mellert, V. Perceptual evaluation of individualized binaural reproduction using a virtual artificial head. *J. Audio Eng. Soc.*, 65(6), pp. 448-459, 2017.

[6] Blau, M., Budnik, A., van de Par, S. Assessment of perceptual attributes of classroom acoustics: Real versus simulated room. *Proceedings of the Institute of Acoustics*, Vol.40. Pt.3., 2018.

[7] Rasumow, E., Hansen, M., van de Par, S., Püschel, D., Mellert, V., Doclo, S., Blau, M. Regularization approaches for synthesizing HRTF directivity patterns. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2), pp. 215-225, 2016.

[8] Fallahi, M., Hansen, M., Doclo, S., van de Par, S., Püschel, D., Blau, M. Individual binaural reproduction of music recordings using a virtual artificial head. *AES Conference on Spatial Reproduction, Tokyo, Japan*, 6−9 August, 2018.

[9] Majdak, P., Balazs, P., Laback, B. Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions. *J. Audio Eng. Soc.*, 55(7/8), pp. 623-637, 2007.