Binaural Reproduction of Signals captured in a reverberant Room with a Virtual Artificial Head

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] *Jade Hochschule Oldenburg, Institüt für Hörtechnik und Audiologie*

[2] *Carl von Ossietzky Universität Oldenburg, Dept. für medizinische Physik und Akustik*

[3] *Akustik Technologie Göttingen* , [4] *Exzellenzcluster Hearing4All*

Introduction

As an alternative to the traditional artificial heads, a microphone array-based filter-and-sum beamformer, referred to as Virtual Artificial Head (VAH) can be used to synthesize the directivity pattern of individual Head Related Transfer Functions (HRTFs) [1]-[2]. The main advantage of the VAH is the possibility to individualize the recordings post hoc by applying the individually calculated filter coefficients to the microphone signals. In addition, head orientation can be changed retrospectively to allow for head tracking. These filter coefficients can be calculated by minimizing a least-squares cost function in a constrained optimization problem. This was previously tested for simulated microphone arrays in free-field conditions [3]. The present study investigated a new scenario with real recordings with the VAH in a reverberant auditorium, where array robustness and the presence of reflections from directions other than the direct sound of the source play a role. Additionally, head tracking was employed during binaural reproduction, a feature which is another advantage of a VAH. The results were evaluated and discussed based on subjective perceptual outcomes of experiments with dynamic binaural presentations.

VAH as beamformer with constraints on WNG and Spectral Distortion

The VAH as a filter-and-sum beamformer aims at synthesizing individual HRTFs (left and right) with directivity pattern $D(f,\Theta_k)$. f denotes the frequency and Θ_k, $k = 1,2,...,P$, denotes the direction. The synthesized directivity pattern $H(f,\Theta_k)$ of this beamformer at direction Θ_k and frequency f is defined as

$$H(f,\Theta_k) = \mathbf{w}^H(f)\mathbf{d}(f,\Theta_k). \quad (1)$$

The $N \times 1$ steering vector $\mathbf{d}(f,\Theta_k)$ describes the free-field acoustic transfer functions between the source at direction Θ_k ($k = 1,2,...,P$) and the N microphones in the array. The $N \times 1$ complex-valued vector $\mathbf{w}(f)$ contains the filter coefficients (FCs) for the N microphones. To calculate these FCs, a narrow-band least-squares cost function

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

was minimized. In order to increase the robustness of the microphone array against deviations in microphone characteristics, positions, and microphone self-noise, the minimization of J_{LS} was performed subject to a constraint on the resulting *mean* White Noise Gain, WNG_m, which is defined as the ratio between the mean output power of the beamformer over all P directions and the output power for spatially uncorrelated white noise [1], i.e.

$$WNG_m = 10\lg\left(\frac{1}{P}\sum_{k=1}^{P}\frac{|\mathbf{w}^H(f)\mathbf{d}(f,\Theta_k)|^2}{\mathbf{w}^H(f)\mathbf{w}(f)}\right)\text{dB} \geq \beta, \quad (3)$$

with β denoting a minimum desired value for the WNG_m. In order to achieve a small synthesis error at all P directions, additional constraints were imposed on the Spectral Distortion (SD) at all directions Θ_k, by setting an upper and lower limit, L_{Up} and L_{Low}, i.e. for all k

$$L_{\text{Low}} \leq SD(f,\Theta_k) = 10\lg\frac{|\mathbf{w}^H(f)\mathbf{d}(f,\Theta_k)|^2}{|D(f,\Theta_k)|^2}\text{dB} \leq L_{\text{Up}}. \quad (4)$$

The minimization of J_{LS} subject to inequality constraints in Eq.(3) and (4) was done using an Interior-Point optimization algorithm with solutions proposed in [1] as the initial values for the iterative method.

The performance of the VAH depends on a variety of parameters such as the constraint parameters β, L_{Up} and L_{Low}, the number P of the directions included in the calculation of the FCs, and the microphone array topology. Applying different P and different constraint values to two simulated microphone arrays of different topologies, it was shown in [3] that with properly chosen array topology and constraint parameters, the VAH can perceptually outperform a classical artificial head with respect to overall audio quality for music content, for source directions in the horizontal plane. However, these results were achieved only with simulated microphone arrays in free-field conditions. In addition, the simulated

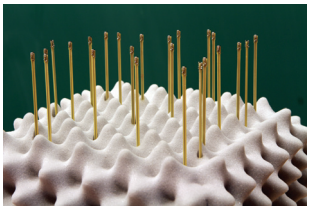

Figure 1: Virtual Artificial Head (VAH) used in this study: planar microphone array with 24 microphones [1].

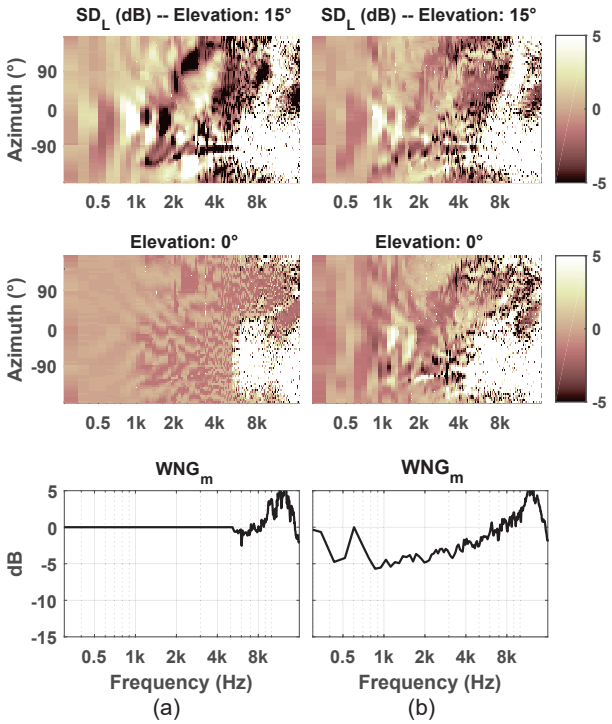

Figure 2: Resulting Spectral Distortion (SD) for synthesized left HRTFs (SD_L) at 15° and 0° elevations as well as the resulting left WNG_m, using filter coefficients calculated with: (a) $P = 72$ directions in the horizontal plane and $\beta = 0$ dB, (b): $P = 3\times\ 72 = 216$ directions from elevations -15°, 0°, 15° and $\beta = 0$ dB.

microphone arrays were considered as completely noise-free and robust, such that the effect of different values for the resulting WNG_m could not be assessed. Therefore, the present study considered a new scenario which consisted of real measurements with the VAH in a reverberant environment. The VAH shown in Fig. 1, i.e. a planar microphone array of 24 microphones, was used for the measurements. The performance of the VAH was evaluated perceptually within dynamic binaural signal representations, allowing for head rotations during headphone presentations.

For this new scenario, it was decided to set the two constraint parameters L_{Low} and L_{Up} to the fixed values of -1.5 dB and 0.5 dB, respectively, leading to a desired maximum deviation of 2 dB of the Interaural Level Differences at all P directions. The two other parameters β and P were modified to examine the performance of the VAH. The extent to which the resulting SD would remain between the two limits of -1.5 dB and 0.5 dB depends on the values chosen for β and P as the following example depicts: Fig. 2a shows the resulting WNG_m and the SD for synthesized left HRTFs for FCs for the left ear calculated with $\beta = 0$ dB and $P = 72$ directions from the horizontal plane (5° azimuthal resolution). While a minimum value of 0 dB for the WNG_m and -1.5 dB $\leq SD \leq$ 0.5 dB for synthesized HRTFs at the 72 directions in the horizontal plane were achieved for frequencies up to 6 kHz, the resulting SD increased clearly at other directions such as at 15° elevation. In a second case, $P = 216$ direction (3×72) from elevations $-15°$, 0° and 15° were considered to calculate the FCs. As Fig. 2b shows, by including the directions from 15° elevation, the resulting SD at this elevation improved compared to the first case, while the results got worse for the synthesized HRTFs in the horizontal plane and for the resulting WNG_m, indicating the impact of P on the resulting SD and WNG_m. In this study, three cases for P were considered: $P = 72$ horizontal directions, $P = 3 \times 72 = 216$ directions from elevations $-15°$, 0° and 15°, and $P = 3 \times 72 = 216$ directions from elevations $-30°$, 0° and 30° (**labeled as El0, El0±15 and El0±30, respectively**, in the remaining discussion). For β, the two cases $\beta = 0$ dB and $\beta = -10$ dB (**labeled as β_0 and β_{-10}**) were considered, which, in combination with the three cases for P, resulted in a total of six sets of FCs.

In order to enable a dynamic binaural signal representation with head rotations, each of the six sets of FCs were calculated for 185 head orientations (azimuth angles $-90°$ to $+90°$ in 5° steps and elevations $-15°$ to $+15°$ in 7.5° steps). FCs for a given head orientation Θ_m, $m \in 1, 2, ..., P$, can simply be calculated by taking the $D(f, \Theta_k)$, $k = 1, 2, ..., P$, and the shifted steering vectors $\mathbf{d}(f, \Theta_{k'})$ with $k' = m, m+1, ..., P, 1, 2, ...m-1$ into Eqs.(1) to (4).

Recordings with the VAH in a reverberant room

The room in this study was a small lecture room (7.12m × 11.94m × 2.98m) with six rows of tables and chairs, and with an average reverberation time of 0.58 s. A listener position was selected in the third row with ears at 1.30m height from the floor. Four source positions were considered in the room (see Fig. 3) : Source 1 (Genelec type 8030c) was located ahead of the listener, slightly higher than the ears to represent the lecturer. Source 2 and Source 3 (Genelec type 8030b) were placed left and right behind the listener, respectively, both at the same height as the ears, representing other speakers in the room. Source 4 was a permanently installed loudspeaker in the room (Event active studio monitor), mounted in front of the room on the right and at an elevation of about 20°. In the next step, the VAH was positioned at the listener position and Room Impulse Responses (RIRs) were measured for each of the 24 microphones of the array and for each source. These RIRs were then filtered with the previously mentioned six sets of FCs to result in six sets of individually synthesized Binaural Room Impulse Responses (referred to as VAH BRIRs), each for 185 head orientations. The RIR measurement was also performed with the KEMAR artificial head and a head-sized rigid sphere (radius = 8.5cm) with two microphones positioned at ±100° on the equator of the sphere (referred to as KEMAR BRIRs and Sphere BRIRs, respectively). In order to enable a dynamic binaural representation with KEMAR and Sphere BRIRs, at least for head orientations in the horizontal plane, the RIRs had to be measured 37 times for 37 orientations of the KEMAR or the rigid

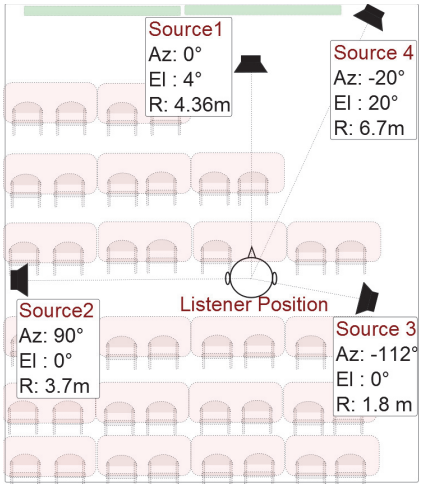

Figure 3: Listener position as well as four source positions for the measurements in the lecture room (Az: azimuth, El: elevation, R: distance to the source)

sphere (azimuthal orientations -90° to +90°, in 5° steps).

Perceptual evaluation and discussion

In order to evaluate the quality of the six VAH BRIRs as well as the KEMAR and Sphere BRIRs, a listening test with dynamic binaural presentations was performed. The headphone was equipped with the tracker on its top bow (see [4] for more details). A push button enabled switching from headphone to loudspeaker presentations, as also implemented in [4]. The listening test took place in the same room and for the same source-listener positions as for the measurements. A total of 4 normal-hearing subjects took part in the test. For all of them, individually measured HRTFs and Headphone Transfer Functions (HPTFs), as well as six sets of individually calculated FCs for 185 head orientations were available. Subjects sat at the listener position during the test and were asked to rate eight different headphone presentations, generated either with VAH BRIRs or with KEMAR and Sphere BRIRs, compared to the reference signal which was real loudspeaker playback in the room. The evaluation was performed three times for each source in a randomized order. The test signal was a dry recorded speech utterance of 15 s duration, spoken by a female speaker, which was convolved with different BRIRs and inverse individual HPTFs for the headphone presentation. Five attributes were considered to be evaluated: Reverberance, Source Width, Source Distance, Source Direction and Overall Quality. Subjects gave their ratings on a 9-point scale covering the five german labels schlecht (bad), dürftig (poor), ordentlich (fair), gut (good) and ausgezeichnet (excellent) and four equidistant intermediate categories. The test signal was presented in a continuous loop and subjects could switch freely between the eight different headphone presentations or between headphone and loudspeaker presentation.

There was no dependency of source directions in any of

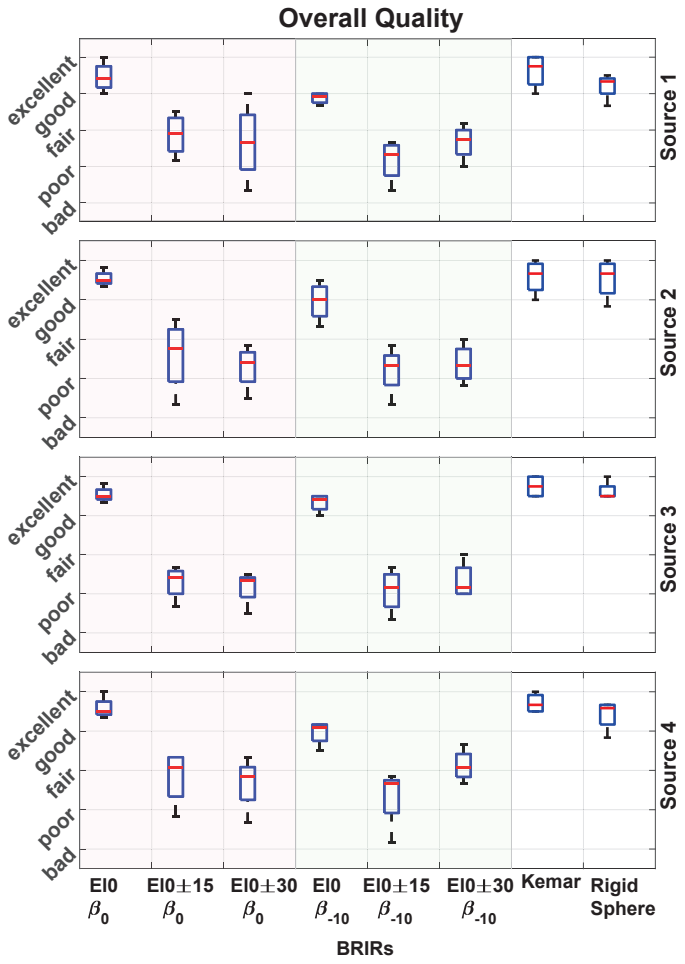

Figure 4: Results of perceptual evaluations for 4 subjects with respect to Overall Quality for different source positions, regarding different BRIRs.

the evaluated perceptual attributes. Exemplarily, Fig. 4 shows the ratings of Overall Quality for all source directions. In the remaining discussion, the ratings for the four source directions and three repetitions were pooled separately for each perceptual attribute, with results shown in Fig. 5. For all attributes, the VAH BRIRs with El0(β_0) and the KEMAR and Sphere BRIRs were rated similarly high, with median ratings between good and excellent. Including more directions in the calculation of FCs in El0±15 and El0±30 led to poorer synthesis accuracy in the horizontal plane, as already discussed in Fig. 2. This matches with the generally lower ratings given to these BRIRs. These results also show the importance of the accuracy in the horizontal plane for sources in and near the horizontal plane. The slightly lower ratings for VAH BRIRs with El0(β_{-10}) compared to BRIRs with El0(β_0) could be due to reduced robustness of the VAH. For FCs with El0(β_{-10}), the constraint on the resulting WNG_m was relaxed with allowable values down to −10 dB, which increases the sensitivity of the VAH synthesis to deviations in microphone characteristics. Taking into account the time lapse of about four months between the measurement of steering vectors for the used VAH and the recordings in the room,

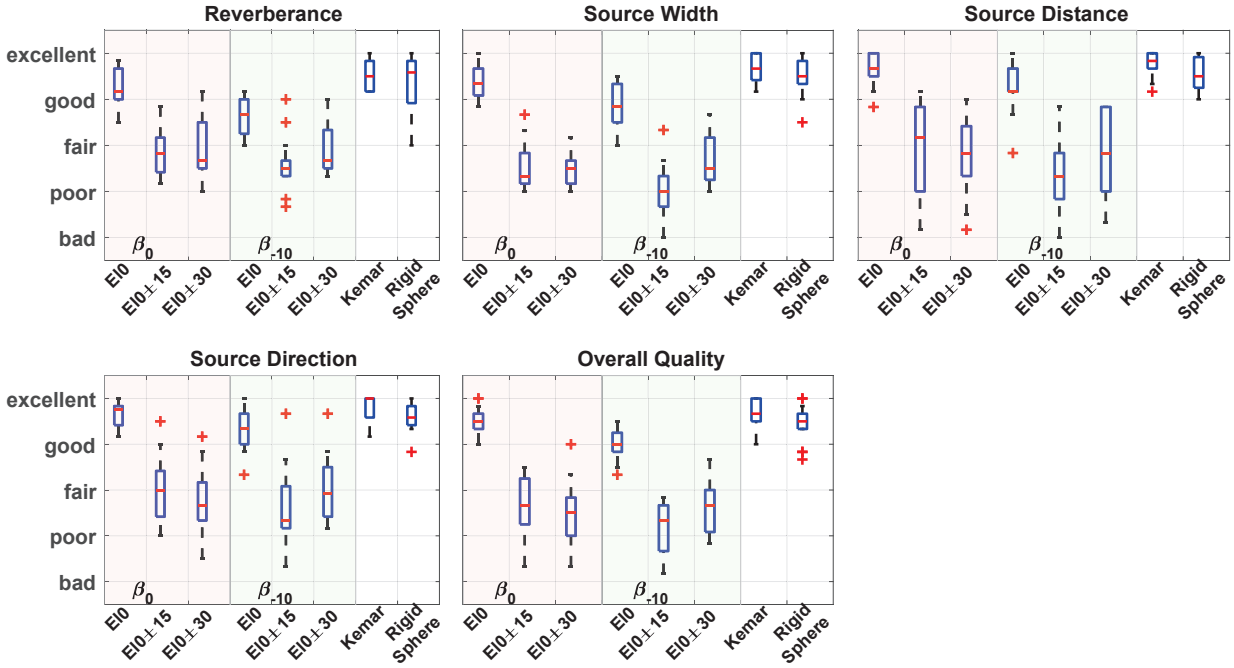

Figure 5: Perceptual ratings averaged across 4 subjects, three presentations, and four source positions regarding different BRIRs.

it seems possible that small unavoidable deviations in microphone positions or characteristics might have occurred during this time period, which caused perceptual artifacts in the synthesized BRIRs. Therefore, it is advisable to choose higher values for β. An interesting point to be discussed is the high ratings given almost everywhere to KEMAR and Sphere BRIRs. One reason might have been the presence of reflections which improved the externalization and helped mask the deficiencies caused by deviations from individual HRTFs. Another important point is that, when supplied with visual information about the room and the loudspeaker positions, it is less likely to have problems such as in-head localization or front-back confusions. Despite the high ratings given to KEMAR or Sphere BRIRs, regarding the impractical effort taken to measure these BRIRs for many different head orientations, and considering the comparable perceptual results achieved with the VAH (VAH BRIRs with El0(β_0)), it seems that for speech stimuli and in a realistic acoustical situation as in this study, the VAH offers the more promising alternative for dynamic binaural auralizations.

Conclusion

In this study, individual BRIRs in a normal lecture room were synthesized with the VAH for different head orientations. It was shown that a typical reverberant room can be dynamically auralized with the VAH for speech signals, with a high perceptual agreement to real loudspeaker presentations in the room. Further investigations should concern the evaluation of the VAH in other environments using other signals (e.g. music) and for other source positions in room, as well as the enhancement of the microphone array topology.

Acknowledgement

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

References

[1] 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

[2] Atkins, J.: Robust beamforming and steering of arbitrary beam patterns using spherical arrays. *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA), New Paltz, NY, USA, Oct. 2011*, pp. 237-240

[3] 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, Aug. 2018*

[4] Blau, M., Budnik, A., van de Par, S.: Assessment of perceptual attributes of classroom acoustics: Real versus simulated room. *Proc. of Institute of Acoustics, Vol.40.Pt.3.2018*