Audio Engineering Society

Convention Paper 9944

Presented at the 144th Convention
2018 May 23 – 26, Milan, Italy

BRIR synthesis using first-order microphone arrays

Markus Zaunschirm, Matthias Frank, and Franz Zotter

Institute of Electronic Music and Acoustics, University of Music and performing Arts, Graz

Correspondence should be addressed to Markus Zaunschirm (zaunschirm@iem.at)

ABSTRACT

Both the quality and immersion of binaural auralization benefit from head movements and individual measurements. However, measurements of binaural room impulse responses (BRIRs) for various head rotations are both time consuming and costly. Hence for efficient BRIR synthesis, a separate measurement of the listener-dependent part (head-related impulse responses, HRIR) and the room-dependent part (RIR) is desirable. The room-dependent part can be measured with compact first-order microphone arrays, however the inherent spatial resolution is often not satisfying. Our contribution presents an approach to enhance the spatial resolution using the spatial decomposition method in order to synthesize high-resolution BRIRs that facilitate easy application of arbitrary HRIRs and incorporation of head movements. Finally, the synthesized BRIRs are compared to measured BRIRs.

1 Introduction

The binaural auralization of acoustic environments or the virtualization of an acoustic scene is typically achieved by convolving a source signal with measured binaural room impulse responses (BRIRs) (individual or using an artificial head) and a subsequent playback over headphones [1, 2]. Typically, the BRIR measurements are both time consuming and costly as the artificial head or each future listener has to be carried to the room to be measured. Moreover, BRIRs have to be measured for various head-rotations in order to allow for dynamic binaural rendering.

Dynamic binaural rendering of object-based audio with BRIRs typically requires a switching of the IRs [3], while rendering using a BRIR represented in Ambisonics (scene-based audio) allows for simple rotation by a frequency-independent matrix multiplication [4, 5] while keeping the IRs of the binaural renderer static.

An efficient and versatile method for BRIR synthesis requires a separation in a listener-dependent and a room-dependent part [6, 7]. The listener-dependent part is typically described by high-resolution far-field head-related impulse responses (HRIRs), and the room-dependent part contains the spatio-temporal information at the listening position. The most efficient way (little hardware effort) to capture the room-dependent part employs a first-order microphone array. However, it has been shown in [8] that the first-order representation of RIRs is not sufficient to preserve decorrelation in the reverberation and results in decreased perceived spatial depth for loudspeaker playback.

Higher directional resolution can be achieved by higher-order microphone arrays (e.g. mh acoustics eigenmike) or by directional sharpening of the first-order RIRs using the spatial decomposition method (SDM) [9].

In this method, directional sharpening is achieved by assigning a discrete direction to each sample of the omnidirectional RIR, where the directions are estimated using e.g. the pseudo-intensity vector (PIV) method [10]. SDM allows for a re-encoding of the measured RIR to any desired Ambisonics order. However, directional sharpening of the RIRs leads to an unnatural increase of the reverberation time at high frequencies, especially when using high encoding orders and thus, an order-dependent spectral correction is necessary [8, 11].
In this paper we present an efficient method for synthesizing BRIRs using measured first-order/4-channel (tetrahedral microphone array, see Fig. 1(b)) RIRs followed by directional sharpening and a convolution with pre-measured high-resolution HRIRs of an artificial head [12]. In a listening experiment, the synthesized BRIRs are compared to BRIRs, which were measured with the same artificial head (KU 100, see Fig. 1(a)). Experiments are conducted for different rooms (different reverberation times), various source positions, and evaluate the perceptual attributes of source width, source distance and diffuseness. The tested BRIR synthesis methods include measured first-order RIRs, and directionally sharpened RIRs using different Ambisonics orders, as well as mapping to the nearest direction available in the HRIR grid.

(a) Neumann *KU 100* artificial head.

(b) TSL *ST450* 4-channel Ambisonic B-format microphone array.

Fig. 1: Measurement equipment used.

2 Measurement-Based BRIR Synthesis

There are various approaches for obtaining BRIRs from measured RIRs of a compact, first-order spherical microphone array, see Fig. 2. The two-staged process consists of (i) extraction of directional information and recombination in the *Directional RIR* stage followed by a (ii) combination of the room-dependent and the listener-dependent part (HRIRs) in the *Rendering stage*.

2.1 Directional RIR

For efficient and low-effort measurements we suggest using a compact first-order tetrahedral spherical microphone array, albeit any 3D array configuration can be used. We define the discrete-time single-input-multiple-output (SIMO) RIRs $h(t)$, $x(t)$, $y(t)$, $z(t)$ as the responses of the four-channel output of the *ST450* array (see Fig. 1(b)) after deconvolution by the measurement signal. The four channels (B-format) correspond to a RIR measurement with four independent directivity patterns: omnidirectional for $h(t)$, figure-of-eight in x for $x(t)$, y for $y(t)$, and z-direction for $z(t)$. Similar to the SDM approach [9], we assign a DOA $\boldsymbol{\theta}(t)$ to each discrete-time sample t of the RIR $h(t)$.

For the DOA estimation, we suggest using the pseudo-intensity vector (PIV) approach in the frequency range from 200 Hz and 3 kHz where the directivity patterns of the microphone can be regarded as coincident and clean [10]. We perform a zero-phase band limitation (e.g. by MATLAB's filtfilt) denoted by $F_{200-3\text{k}}$ and a zero-phase smoothing F_L of the resulting PIV using a moving-average time window on the interval $[-L/2;L/2]$ for $L = 16$ around each sample and get the DOA estimate

$$\boldsymbol{\theta}(t) = \frac{\tilde{\boldsymbol{\theta}}(t)}{\|\tilde{\boldsymbol{\theta}}(t)\|}, \quad \text{with} \tag{1}$$

$$\tilde{\boldsymbol{\theta}}(t) = F_L\left\{ F_{200-3k}\left\{h(t)\right\} F_{200-3k}\left\{\begin{bmatrix}x(t)\\ y(t)\\ z(t)\end{bmatrix}\right\}\right\}$$

$\boldsymbol{\theta}(t)$ as Cartesian unit vector.

2.2 Rendering Methods

In order to obtain the synthesized BRIRs, the *directional RIR* is combined with the listener-dependent part, the HRIRs. Now let us consider an arbitrary HRIR set

$$\boldsymbol{A}(t) = [\boldsymbol{a}_1(t), \cdots, \boldsymbol{a}_p(t), \cdots, \boldsymbol{a}_P(t)], \tag{2}$$

$$\boldsymbol{a}_p(t) = [a_p^l(t), a_p^r(t)]^T, \tag{3}$$

where $(\cdot)^{l,r}$ indicates the left and right ear, $(\cdot)^T$ is the transpose operator, the index p indicates the $p-$th direction defined by the normalized Cartesian direction vector $\boldsymbol{\theta}_p = [x_p, y_p, z_p]^T$ of the HRIR sampling grid, and P is the total number of HRIRs.

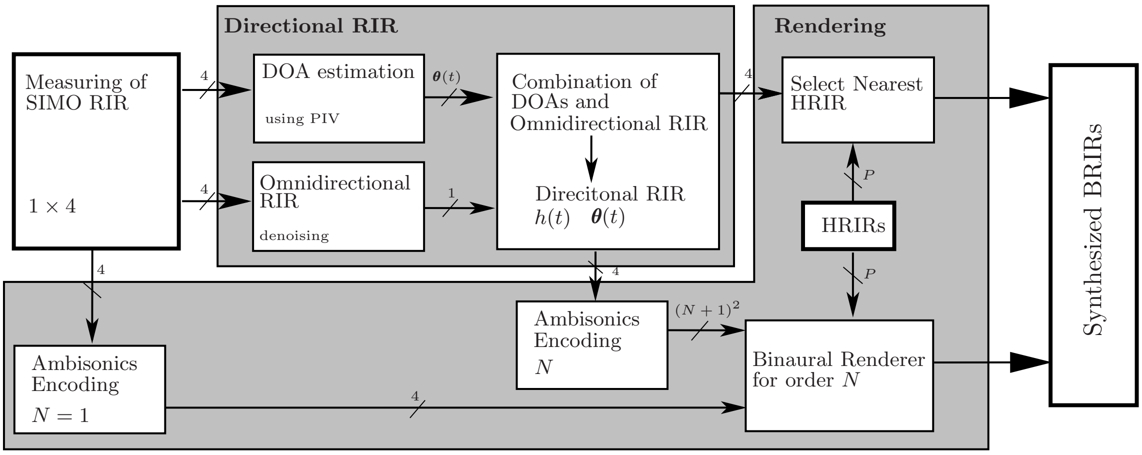


Fig. 2: Block diagram of BRIR synthesis. The measured SIMO RIRs are used for extracting an omnidirectional RIR and corresponding DOA estimates at the receiver location. In the rendering stage the directional RIR is either represented in the Ambisonics domain (see Eq. (8)) and rendered via a state-of-the-art Ambisonics renderer or directly rendered by selecting a HRIR from a pre-measured data set, see Eq. (4).

The synthesized BRIRs are either obtained by (i) time-variant selection of the HRIR direction nearest to the estimated DOA $\boldsymbol{\theta}(t)$, or (ii) binaural rendering of the directional RIR represented in Ambisonics.

2.2.1 Nearest Neighbor Selection

With the unit vector $\boldsymbol{\theta}(t)$, the BRIR synthesis by selection of the nearest-neighbor HRIR pair $a_p^{l,r}(t)$ is

$$BRIR^{l,r}_{NN_P}(t) = \sum_{\tau=0}^{T-1} h(\tau)\, a^{l,r}_{\tilde{p}(\tau)}(t-\tau), \qquad (4)$$

$$\tilde{p}(t) = \arg\max_{p} \; \boldsymbol{\theta}_p^T \boldsymbol{\theta}(t), \qquad (5)$$

where $\tilde{p}(t)$ corresponds to the HRIR sampling grid index that is closest to the DOA estimation at discrete time index t, and T is the length of $h(t)$.

2.2.2 Ambisonics

The omnidirectional impulse response $h(t)$ is mixed using the time-dependent direction-of-arrival vector $\boldsymbol{\theta}(t)$ to get a first version of a higher-order Ambisonics room impulse response

$$\tilde{h}_{nm}(t) = Y_n^m[\boldsymbol{\theta}(t)]\, h(t), \qquad (6)$$

where $Y_n^m(\boldsymbol{\theta})$ are the N3D-normalized, real-valued spherical harmonics of order n and degree m evaluated at the direction $\boldsymbol{\theta}$, and N is the maximum order.

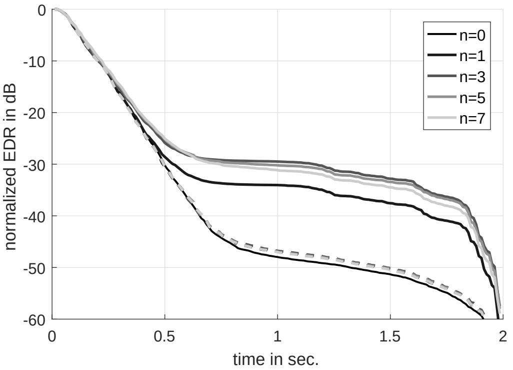


Fig. 3: Energy decay relief (EDR) in a third-octave band with center frequency of 2 kHz. Solid and dashed lines indicate the order partitioned EDR before and after equalization as defined in Eqs. (6) and (7), respectively.

A fast variation of the direction of arrival $\boldsymbol{\theta}(t)$ causes strong amplitude modulation and destroys narrow-band spectral content in $\tilde{h}_{nm}(t)$ by spectral whitening; typically, the longer low-frequency reverberation tails are hereby mixed towards higher frequencies, causing unnaturally long reverberation there [8, 11], cf. solid lines in Fig. 3. Therefore, the response $\tilde{h}_{nm}(t)$ needs spectral correction. To this end, third-octave filtering $\tilde{h}_{nm}(t) = \sum_b F_b\{\tilde{h}_{nm}(t)\}$ is useful, where the b-th subband signal with center frequency f_b is obtained by perfectly reconstructing zero-phase filters F_b.

The time-variant envelope $w_n^b(t)$ accomplishes spectral correction of the sub-band response

$$F_b\{h_{nm}(t)\} = F_b\{\tilde{h}_{nm}(t)\}\, w_n^b(t), \qquad (7)$$
$$\text{with } w_n^b(t) = \sqrt{\tfrac{2n+1}{4\pi}}\sqrt{\frac{F_T\{F_b\{h(t)\}^2\}}{\sum_m F_T\{F_b\{h_{nm}(t)\}^2\}}},$$

and an averaging time T (e.g. 10 ms), as derived in appendix A. Dashed lines in Fig. 3 show corrected energy decays of higher-order Ambisonic RIRs for the third octave $f_b = 2$ kHz and the orders $n = \{1,3,5,7\}$.

From $h_{nm}(t) = \sum_b F_b\{h_{nm}(t)\}$ the BRIRs for an order N are synthesized by

$$BRIR^{l,r}_{A_N}(t) = \sum_{\tau=0}^{T-1}\sum_{n=0}^{N}\sum_{m=-n}^{n} b^{l,r}_{nm}(\tau)h_{nm}(t-\tau), \qquad (8)$$

where $b^{l,r}_{nm}(t)$ is any state-of-the-art FIR binaural Ambisonic renderer (e.g. [13]) of the length T, or the one favored here that was defined in [14]. Its frequency-dependent time-alignment of the HRIR set and a diffuse-field constraint can significantly improve both the coloration as well as localization accuracy of binaurally rendered Ambisonic signals represented by practical orders $N < 7$.

3 Evaluation

The proposed BRIR synthesis methods are compared and evaluated via both technical measures including the reverberation time (T30), early decay time (EDR), clarity index (C80), apparent source width (ASW) defined as $1 - IACC_E$, and a listening experiment against a reference BRIR.

The reference BRIRs are recorded in real rooms between a single *Genelec 8020* loudspeaker and the *KU 100* artificial head using the exponentially swept sine method [15]. All SIMO RIRs of the *ST450*, which are the basis of the synthesized BRIRs (see Fig. 2), are measured with the same source, at the same position, and using the same excitation signal.

Overall the test conditions include (i) three different rooms at the IEM Graz, and (ii) two different directions or source and receiver distances at a fixed source- and ear-height of 1.3m. In all measurements the source is directed towards the artificial head/microphone array.

The measured rooms, directions ϕ (azimuth angle with origin in center of the artificial head, and positive x-axis through the nose) and source-receiver distance r are

- Production studio (PS): volume $127\,\mathrm{m}^3$, base area $42\,\mathrm{m}^2$, $\mathrm{T}_{60} \approx 0.4\,\mathrm{s}$. Directions: $\phi = 0°$, and $\phi = 90°$. Source distance $r = 2.3\,\mathrm{m}$.
- CUBE (CU): volume $620\,\mathrm{m}^3$, base area $130\,\mathrm{m}^2$, $\mathrm{T}_{60} \approx 0.7\,\mathrm{s}$. Directions: $\phi = 0°$, and $\phi = 90°$. Source distances $r = 2.3\,\mathrm{m}$, and $r = 4\,\mathrm{m}$.
- Corridor (CO): volume $210\,\mathrm{m}^3$, base area $64\,\mathrm{m}^2$, $\mathrm{T}_{60} \approx 1.4\,\mathrm{s}$. Direction $\phi = 0°$. Source distance $r = 10\,\mathrm{m}$.

For synthesis we used the omni-directional (W-channel) of the *ST450* output and for rendering a far-field HRIR data set of the *KU 100* measured at overall $P = 2702$ sampling points [12]. The tested synthesis methods include

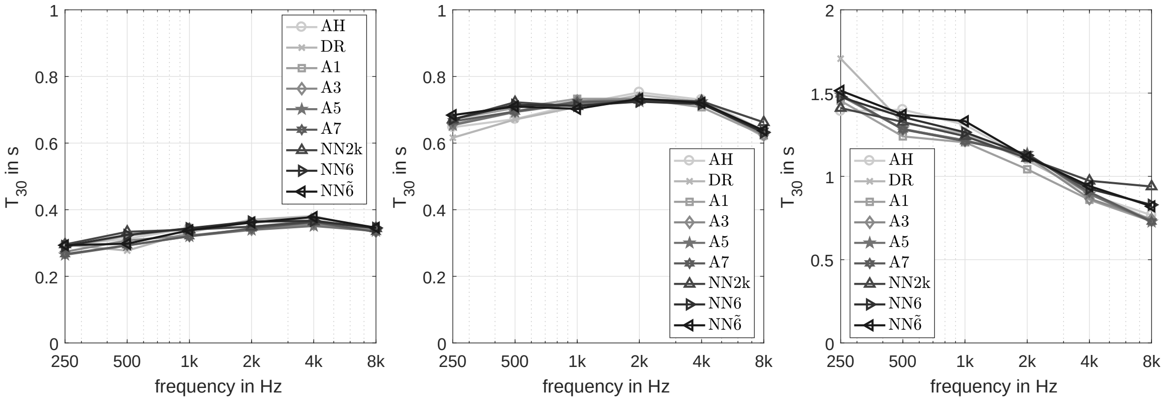
- Direct rendering (DR) of the measured B-format RIRs, see Eq. (8).
- Nearest neighbor rendering with the entire HRIR set (NN2k), with a subset consisting of 6 HRIRs at the front, left, back, right, top, and bottom (NN6), and a subset of 6 HRIRs at front-left, back-left, back-right, front-right, top, and bottom ($\mathrm{NN}\tilde{6}$), see Eq. (4).
- Rendering using the directionally sharpened RIR with orders $N = \{1,3,5,7\}$. The corresponding synthesized BRIRs are abbreviated as A1, A3, A5, A7, respectively.

3.1 Technical Measures

In the following section, the measured reference and synthesized BRIRs are analyzed in terms of technical measures as defined in [16]. For all measures which require a single channel input, the BRIRs of the left and right ear are averaged and parameters are calculated with the Lundeby method [17] which is employed in the AcMus toolbox [18]. As the RIRs are measured for multiple directions and source distances, the parameters are averaged to give a single value per room.

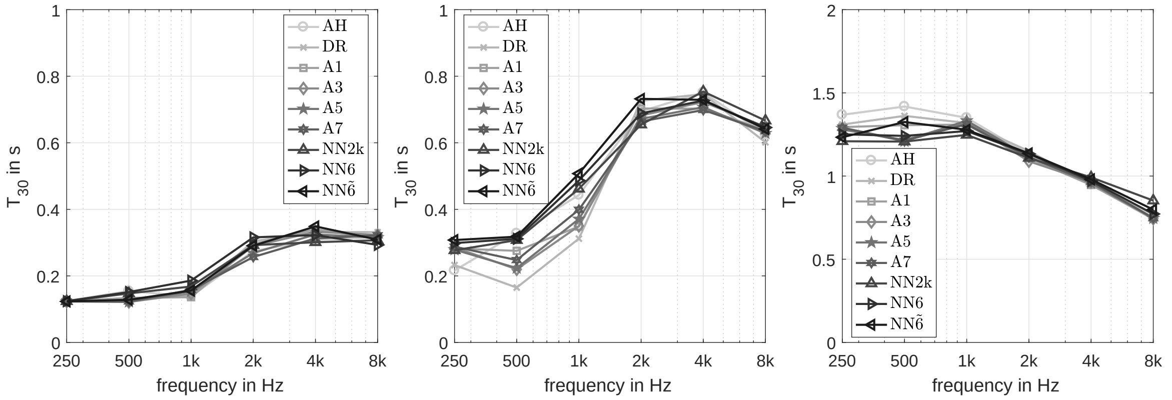
3.1.1 Reverberation Time

The typical measure for the energy decay rate in a room is the reverberation time, which is typically calculated via the Schroeder backwards integration in octave bands between 250 Hz and 8 kHz [16]. The resulting T_{30} values for all synthesis methods and for each of the measured rooms are depicted in Fig. 4. As expected, little variation is observed across the rendering methods as the processing not alters the energy decay of the measured omnidirectional response.



(a) Production Studio. (b) CUBE. (c) Corridor.

Fig. 4: Reverberation time T_{30} in octave bands between 250 Hz and 8 kHz for the three evaluated rooms.



(a) Production Studio. (b) CUBE. (c) Corridor.

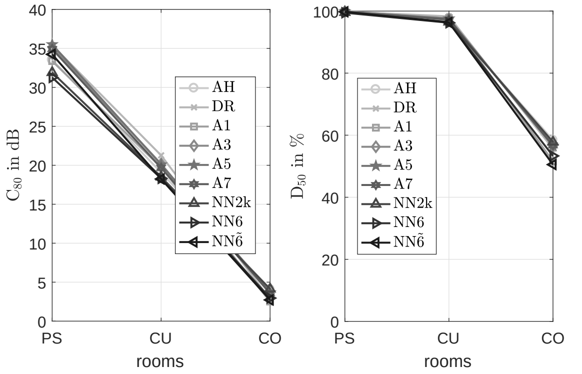
Fig. 5: Early decay time (EDT) in octave bands between 250 Hz and 8 kHz for the three evaluated rooms.

3.1.2 Early Decay Time

According to [16] the early decay time (EDT) is considered to be a more appropriate technical measure for the reverberance of a room than the reverberation time. As the EDT is based on the slope of the 10dB drop in the normalized energy decay curve (EDC), early reflections contribute more significantly to the EDT when compared to the reverberation time. While the EDTs for PS and CO are well aligned with the reference (AH), a deviation can be observed for CU at the lower octave bands, see Fig. 5. In [16] the JND for EDT is quantified as 5%, however it is pointed out in [19] that the JND is highly dependent on the source signal and that JNDs between 25% can be expected.

3.1.3 Clarity and Definition

The technical measures of speech intelligibility and transparency of music as defined in [16] are the definition (D_{50}) and clarity (C_{80}), respectively. Results for the reference and synthesized BRIRs are shown in Fig. 6. For most conditions and rooms the results show little deviation to the reference (2dB clarity change in CU, and 5% definition change in CO). As expected, clarity and definition are higher in rooms with lower T_{30}, cf.. Fig. 4. According to the JNDs given in [16] (1dB for clarity, and 5% for definition) the worst case deviations lie between 1-3 JNDs (NN$\tilde{6}$ in CU).



(a) Clarity index. (b) Definition.

Fig. 6: Single number clarity C_{80} and definition D_{50} for all rooms (averaged between 500 Hz and 1 kHz octave band).

3.1.4 Apparent Source Width

In [20] the apparent source width ASW is defined as *apparent auditory width of the sound field created by a performing entity as perceived by a listener* and modeled via a measure related to the interaural cross-correlation coefficient (IACC). Here the IACC is calculated in the integration interval between [0,80]ms and $(1 - IACC_{[0,80]ms})$ is depicted in Fig. 7. It can be seen that DR, NN6, and NN$\tilde{6}$ show significant deviations from the reference (JND for IACC is defined frequency-independent as 0.075 in [16], although it has been shown in [21] that the JNDs strongly depend on the reference condition and range from $0.08 - 0.35$).

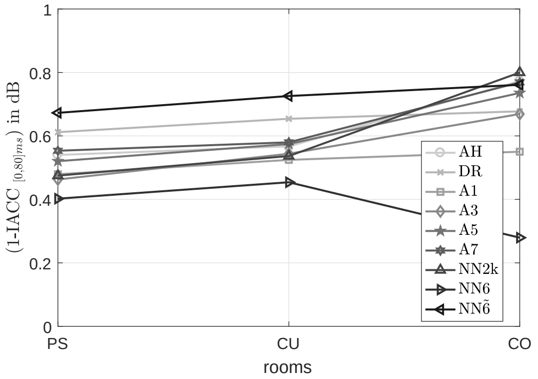


Fig. 7: $(1 - IACC_{[0,80]ms})$ for all rooms and BRIR synthesis methods.

3.2 Listening Experiment

For the common technical measures, which are based on energy decay or ratios, no clear difference is observed between the reference (AH) and synthesized BRIRs, see Sec. 3.1. In order to evaluate audible differences, a listening experiment was conducted. The experiment compared the above-mentioned synthesis/rendering methods (DR, A1, A3, A5, A7, NN2k, NN6, and NN$\tilde{6}$) in a MUSHRA-like [22] procedure against the artificial head AH as a reference.

Note that global timbral deviations between the reference and the synthesized BRIRs, which occur due to spatial aliasing [23], array imperfections (encoding) and microphone frequency responses, are equalized with a single global minimum-phase equalization filter for all synthesis methods and both ears.

The comparison evaluated 3 attributes that seemed reasonable from informal listening by the authors:

- Width: how wide is the source spread and how blurry is the localization of the direct sound?
- Diffuseness: how evenly is the reverberation distributed, are distinct spatial areas audible?
- Distance: how far is the source perceived?

As the localization of the direct sound was only altered by the NN$\tilde{6}$ renderer, this attribute was omitted.

The experiment included all 7 room conditions (PS ($r = 2.3$ m) and CU ($r = \{2.3, 4.0\}$ m) for source directions of $\phi = 0°, 90°$ and CO ($r = 10$ m) for $\phi = 0°$). It was divided into 3 parts, one for each attribute. The parts were performed in random order and within each part, room conditions and rendering methods were also randomized. The source signal were the first 5 seconds of the EBU female German speech recording [24]. Playback employed equalized AKG K702 headphones powered by an RME Multiface.

6 listeners with experience in spatial audio (all male, average age 33 years) participated in the experiment and it took them on average 56 minutes.

3.2.1 Results

For each attribute, room condition, and listener, the answers were normalized to a maximum absolute difference of 1 to the reference. The results for width and diffuseness could be summarized over all 7 room conditions. However for distance, their was a clear separation into two groups: one with frontal direct sound (4 conditions) and one with direction sound from the side (3 conditions).

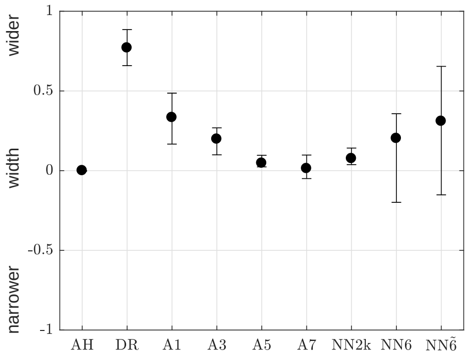


Fig. 8: Median values and corresponding 95% confidence intervals for perceived width, summarizing all 6 listeners and 7 room conditions.

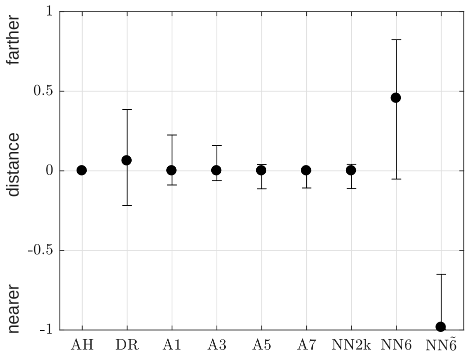


Fig. 9: Median values and corresponding 95% confidence intervals for perceived distance, summarizing all 6 listeners and 4 room conditions with frontal direct sound (PS ($r = 2.3\,\mathrm{m}$), CU ($r = \{2.3, 4.0\}\,\mathrm{m}$), and CO ($r = 10.0\,\mathrm{m}$) for $\phi = 0°$).

Direct rendering of the 1st-order RIR (DR) was perceived significantly wider ($p \leq 0.018$) than all other rendering methods and the reference AH, cf. Fig. 8. Width decreases with the Ambisonics order, so that A5 and A7 ($p > 0.11$) are not distinguishable from AH. Although from the NN renderers only NN2k is significantly wider than the reference AH ($p = 0.026$), NN6 and NN$\tilde{6}$ exhibit an undesirable large spread.

NN6 and NN$\tilde{6}$ yield significantly less even diffuseness than all other renderers and the reference AH ($p \leq 0.015$), whereas both are not significantly different, see Fig. 10. All remaining renderers are not distinguishable from AH ($p > 0.12$) and among themselves ($p > 0.065$).

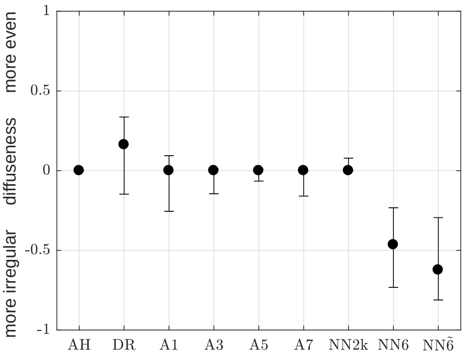


Fig. 10: Median values and corresponding 95% confidence intervals for perceived diffusity, summarizing all 6 listeners and 7 room conditions.

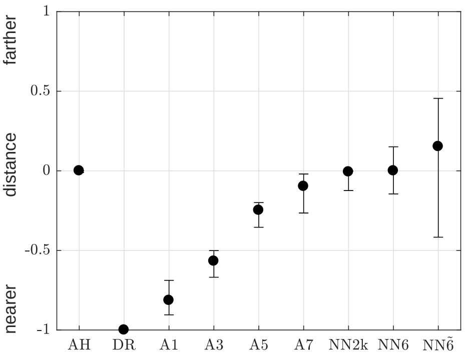


Fig. 11: Median values and corresponding 95% confidence intervals for perceived distance, summarizing all 6 listeners and 3 room conditions with lateral direct sound (PS ($r = 2.3\,\mathrm{m}$), CU ($r = \{2.3, 4.0\}\,\mathrm{m}$) for $\phi = 90°$).

As Fig. 9 shows, NN$\tilde{6}$ yields the smallest distance of all renderers ($p << 0.001$) for frontal direct sound. N6 yields farther results ($p \leq 0.016$) than most renderers, except for DR and A1 ($p > 0.067$). DR, all Ambisonic renderers, and NN2k are indistinguishable from the reference AH ($p > 0.15$).

For direct sound from the side, cf. Fig. 11, the perceived distance significantly increases from DR to A1 to A3 to A5 ($p \leq 0.01$). Further increasing of the order to 7 does not significantly increase the distance ($p = 0.071$), however A7 is the only Ambisonic renderer that is indistinguishable from the reference AH ($p = 0.093$). Moreover, all NN renderers are not distinguishable from the reference AH ($p > 0.33$), however NN$\tilde{6}$ again exhibits an undesirable large spread.

3.3 Discussion

While from the technical measures only the IACC indicates differences between the tested renderers, the listening experiment revealed significant differences for all evaluated attributes. Except for diffuseness and distance of frontal sources, the direct binaural rendering of the 1st-order RIR (DR) significantly deviates from the measured BRIR (AH). Unsurprisingly, the deviation decreases with the Ambisonic order of the directionally sharpened RIRs. With an order of 7, the synthesized BRIR is indistinguishable from the measured BRIR. Similarly good results are obtained for NN2k. In most cases, the BRIRs synthesized by the nearest neighbor renderers with only 6 directions (NN6 and NN$\tilde{6}$) largely differ from the measured BRIR and their results have a large spread. Their results also strongly depend on whether the direction of the direct sound coincides with the directions of the selectable HRIRs. Moreover, the sparse mapping to 6 directions also impairs the evenness of the reverberation, resulting in reduced diffuseness.

4 Conclusion

In this contribution we presented an efficient two-staged measurement-based BRIR synthesis method, which allows for subsequent incorporation of arbitrary HRIRs. In the first stage, the measured SIMO RIRs of a compact tetrahedral microphone array are used for both the extraction of the omnidirectional RIR and time-variant DOA estimation, similar to SDM [9]. In a rendering stage the omnidirectional RIR and DOA estimates are either used for (i) nearest neighbor rendering to directions available in the HRIR set or (ii) Ambisonic rendering of a directionally sharpened RIR. Listening experiments compared the synthesized BRIRs against the measured reference BRIRs for 3 rooms with different acoustic characteristics and source positions. Using Ambisonic rendering, an order of 7 yields results that are indistinguishable from the reference in terms of distance, width, and diffuseness. Reduction of the Ambisonic order increases the deviation from the reference. Interestingly, the sharpened 1st-order Ambisonic rendering outperforms the direct rendering of the measured 1st-order RIRs.

The nearest neighbor rendering with $P = 2702$ HRIR directions yields similar results to 7th-order Ambisonics; results for using $P = 6$ directions show strong deviations from the reference and undesirably large spread and are therefore not recommended.

Moreover, the higher-order representation of the directional RIR allows for rotation of the acoustic scene via a frequency-independent matrix multiplication to account for head movements prior to rendering with a static set of filters for BRIR synthesis.

References

[1] Wightman, F. L. and Kistler, D. J., "Headphone simulation of free field listening I: stimulus synthesis," *J. Acoust. Soc. Am.*, 85(1989), pp. 858–867, 1989.

[2] Møller, H., "Fundamentals of binaural technology," *Applied Acoustics*, 36(3-4), pp. 171–218, 1992, ISSN 0003682X, doi:10.1016/0003-682X(92)90046-U.

[3] Engdegard, J., Resch, B., Falch, C., Hellmuth, O., Hilpert, J., Hoelzer, A., Breebaart, J., Koppens, J., Schuijers, E., and Oomen, W., "Spatial Audio Object Coding (SAOC) – The Upcoming MPEG Standard on Parametric Object Based Audio Coding," *124th AES Convention*, 2008.

[4] Jot, J.-M., Larcher, V., and Pernaux, J.-M., "A Comparative Study of 3-D Audio Encoding and Rendering Techniques," *AES 16th International Conference*, pp. 281–300, 1999.

[5] Pinchon, D. and Hoggan, P. E., "Rotation matrices for real spherical harmonics: General rotations of atomic orbitals in space-fixed axes," *Journal of Physics A: Mathematical and Theoretical*, 40(7), pp. 1597–1610, 2007, ISSN 17518113, doi:10.1088/1751-8113/40/7/011.

[6] Pörschmann, C. and Wiefling, S., "Perceptual Aspects of Dynamic Binaural Synthesis based on Measured Omnidirectional Room Impulse Responses," *International Conference on Spatial Audio*, (December 2016), 2015.

[7] Menzer, F., *Binaural Audio Signal Processing Using Interaural Coherence Matching*, Ph.D. thesis, 2010.

[8] Frank, M. and Zotter, F., "Spatial impression and directional resolution in the reproduction of reverberation," in *Proc. DAGA*, pp. 1304–1307, Aachen, 2016.

[9] Tervo, S., Pätynen, J., Kuusinen, A., and Lokki, T., "Spatial decomposition method for room impulse responses," *Journal of the Audio Engineering Society*, 61(1/2), pp. 17–28, 2013.

[10] Jarrett, D. P., Habets, E. A. P., and Naylor, P. A., "3D Source localization in the spherical harmonic domain using a pseudointensity vector," *European Signal Processing Conference*, (April), pp. 442–446, 2010, ISSN 22195491.

[11] Zaunschirm, M., Baumgartner, C., Schörkhuber, C., Frank, M., and Zotter, F., "An Efficient Source-and-Receiver-Directional RIR Measurement Method," in *Fortschritte der Akustik AIA-DAGA 2017*, pp. 1343–1346, Kiel, 2017.

[12] Bernschütz, B., "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," *Fortschritte der Akustik – AIA-DAGA 2013*, pp. 592–595, 2013.

[13] Bernschuetz, B., Vazquez Giner, A., Poerschmann, C., and Arend, J., "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, 100(5), pp. 972–983, 2014, ISSN 16101928, doi:10.3813/AAA.918777.

[14] Zaunschirm, M., Schoerkhuber, C., and Hoeldrich, R., "Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint," *J. Acoust. Soc. Am., submitted*, 2018.

[15] Farina, A., "Simultaneous measurement of impulse response and distortion with a swept-sine technique," *Proc. AES 108th conv, Paris, France*, (I), pp. 1–15, 2000, doi:10.1109/ASPAA.1999.810884.

[16] BS EN ISO 3382-1:2009, "Acoustics - Measurement of room acoustic parameters. Part 1:Performance spaces," *British Standard*, pp. 1 – 26, 2009, ISSN 1098-6596, doi:10.1017/CBO9781107415324.004.

[17] Lundeby, A., Vigran, T. E., Bietz, H., and Vorländer, M., "Uncertainties of measurements in room acoustics," *Acta Acustica united with Acustica*, 81(4), pp. 344–355, 1995.

[18] Queiroz, M., Iazzetta, F., Kon, F., Gomes, M. H. a., Figueiredo, F. L., Masiero, B., Ueda, L. K., Dias, L., Torres, M. H. C., and Thomaz, L. F., "AcMus: an open, integrated platform for room acoustics research," *Journal of the Brazilian Computer Society*, 14(3), pp. 87–103, 2008, ISSN 0104-6500, doi:10.1007/BF03192566.

[19] Meng, Z., Zhao, F., and He, M., "The just noticeable difference of noise length and reverberation perception," *2006 International Symposium on Communications and Information Technologies, ISCIT*, pp. 418–421, 2006, doi:10.1109/ISCIT.2006.339980.

[20] Hidaka, T., "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls," *The Journal of the Acoustical Society of America*, 98(2), pp. 988–1007, 1995, ISSN 00014966, doi:10.1121/1.414451.

[21] Kim, C., Mason, R., and Brookes, T., "Initial investigation of signal capture techniques for objective measurement of spatial impression considering head movement," in *124th AES Convention, Amsterdam, Netherlands*, volume 7331, 2008, ISBN 9781605602950.

[22] International Telecommunication Union, "ITU-R BS.1534-3, Method for the subjective assessment of intermediate quality level of audio systems," *ITU-R Recommendation*, 1534-3, pp. 1534–3, 2015.

[23] Rafaely, B., Weiss, B., and Bachmat, E., "Spatial aliasing in spherical microphone arrays," *IEEE Transactions on Signal Processing*, 55(3), pp. 1003–1010, 2007, ISSN 1053587X, doi:10.1109/TSP.2006.888896.

[24] European Broadcasting Union, "EBU Tech 3253 - Sound Quality Assessment Material recordings for subjective tests," (September), 2008.

A Spectral Correction of Directionally Sharpend RIRs in Ambisonics

The squared impulse response after SDM upmixing is

$$\tilde{h}^2_{nm}(t) = |Y_n^m[\boldsymbol{\theta}(t)]|^2\, h^2(t), \tag{9}$$

and due to the pseudo-allpass property $\sum_{m=-n}^{n} |Y_n^m(\boldsymbol{\theta})|^2 = 2n+1$ of the orthonormal spherical harmonics, $\int_{\mathbb{S}^2} |Y_n^m(\boldsymbol{\theta})|^2 \,\mathrm{d}\boldsymbol{\theta} = 1$, we obtain a relation between energies in the DOA-modulated n^{th}-order and the 0^{th} order impulse response $\tilde{h}^2_{00}(t) = \frac{1}{4\pi} h^2(t)$,

$$\sum_{m=-n}^{n} \tilde{h}^2_{nm}(t) = \tfrac{2n+1}{4\pi}\, h^2(t). \tag{10}$$

For a spectral correction, we observe that this property remains unaffected when summing over T discrete-time instances $t = -\frac{\mathrm{T}}{2}, \dots \frac{\mathrm{T}}{2} - 1$ around the time instant τ

$$\sum_{t=-\frac{\mathrm{T}}{2}}^{\frac{\mathrm{T}}{2}-1} \sum_{m=-n}^{n} \tilde{h}^2_{nm}(t+\tau) = \frac{2n+1}{4\pi} \sum_{t=-\frac{\mathrm{T}}{2}}^{\frac{\mathrm{T}}{2}-1} h^2(t+\tau), \tag{11}$$

hence Parseval's theorem allows to replace summation over squared discrete-time samples by summation over magnitude squared discrete-frequency Fourier coefficients $\sum_{t=-\frac{\mathrm{T}}{2}}^{\frac{\mathrm{T}}{2}-1} x^2(t) = \sum_{k=0}^{\mathrm{T}-1} |X(k)|^2$. Consequently, the above relation for the energy within the order n of the room response also holds in the frequency domain

$$\sum_{m=-n}^{n} \sum_{k=0}^{\mathrm{T}-1} |\tilde{H}_{nm,\tau}(k)|^2 = \frac{2n+1}{4\pi} \sum_{k=0}^{\mathrm{T}-1} |H_\tau(k)|^2, \tag{12}$$

which finally permits to undertake spectral corrections. To correct the spectrally whitened response $\tilde{H}_{nm,\tau}(k)$, we introduce an equalizer $W_{n,\tau}(k)$ and define the spectrally corrected response for a time-offset τ as

$$H_{nm,\tau}(k) = W_{n,\tau}(k)\, \tilde{H}_{nm,\tau}(k). \tag{13}$$

While the equalizer must retain the above equation for the summed energies over k and m, the equalizer should restore the spectral decay of the response $H_\tau(k)$ for all time shifts τ at all discrete frequencies k

$$\sum_{m=-n}^{n} |W_{n,\tau}(k)\, \tilde{H}_{nm,\tau}(k)|^2 = \tfrac{2n+1}{4\pi} |H_\tau(k)|^2, \tag{14}$$

$$|W_{n,\tau}(k)|^2 \sum_{m=-n}^{n} |\tilde{H}_{nm,\tau}(k)|^2 = \tfrac{2n+1}{4\pi} |H_\tau(k)|^2.$$

Thus the spectral decay correction

$$|W_{n,\tau}(k)|^2 = \frac{2n+1}{4\pi} \frac{|H_\tau(k)|^2}{\sum_{m=-n}^{n} |\tilde{H}_{nm,\tau}(k)|^2}. \tag{15}$$

can be applied to the room impulse response as a time-variant filter

$$h_{nm}(t+\tau) = \sum_{k=0}^{\mathrm{T}-1} W_{n,\tau}(k)\, H_{nm,\tau}(k)\, e^{\mathrm{i}\frac{2\pi}{\mathrm{T}}kt}. \tag{16}$$

For smooth results, a third-octave analysis is advised and smoothing of the envelope $W_{n,\tau}(k)$ to a temporal envelope $w_n^b(t)$ that can be applied to the third-octave decomposed impulse response, as above. The envelope $w_n^b(t)$ is obtained from the square-root of the gathered energies $\sqrt{\sum_{k:f_k \in [2^{-1/6};2^{1/6}]f_b} W_\tau^2(k)}$ of the bins around the third-octave center frequencies f_b. These are interpolated over all instants t between the analysis time shifts τ (hop size). In the current implementation, gathering of the energies for the band b employs a $\sin^2$ window from f_{b-1} to f_{b+1} to get smooth transitions between the bands.