Flexible binaural resynthesis of room impulse responses for augmented reality research

Sebastià V. Amengual Garí, W. Owen Brimijoin, Henrik G. Hassager, Philip W. Robinson

FLEXIBLE BINAURAL RESYNTHESIS OF ROOM IMPULSE RESPONSES FOR AUGMENTED REALITY RESEARCH

Sebastià V. Amengual Garí, W. Owen Brimijoin, Henrik G. Hassager, Philip W. Robinson
Facebook Reality Labs
sebastia.amengual@oculus.com

ABSTRACT

A basic building block of audio for Augmented Reality (AR) is the use of virtual sound sources layered on top of real sources present in an environment. In order to perceive these virtual sources as belonging to the natural scene it is important to carefully replicate the room acoustics of the listening space. However, it is unclear to what extent the real and virtual room impulse responses (RIR) need to be matched in order to generate plausible scenes in which virtual sound sources blend seamlessly with real sound sources. This contribution presents an auralization framework that allows binaural rendering, manipulation and reproduction of room acoustics in augmented reality scenarios, in order to get a better understanding of the perceptual relevance of individual room acoustic parameters. Auralizations are generated from measured multichannel room impulse responses (MRIR) parametrized using the Spatial Decomposition Method (SDM). An alternative method to correct the known time-dependent coloration of SDM based auralizations is presented. Instrumental validation shows that re-synthesized binaural room impulse responses (BRIRs) are in close agreement with measured BRIRs. In situ perceptual validation with expert listeners show that - in the presence of visual cues, an explicit sound reference and unlimited listening time, they are able to discriminate between a real loudspeaker and its re-synthesized version. However, the renderings appear to be as plausible as a real source once visual cues are removed. Finally, approaches to manipulate the spatial and time-energy properties of the auralizations are presented.

1. INTRODUCTION

One of the main applications of audio for AR is the integration of virtual objects into the real environment. These could be either augmented versions of real objects, e.g., adding to or changing the sound from real objects - or the creation of fully virtual objects, e.g., human avatars. In both cases, in order to deliver a convincing and coherent experience, the acoustic properties of the virtual and real environment must match. However, multiple elements are involved in this audio rendering pipeline i.e. individualization of head-related transfer functions (HRTFs), sound propagation modeling, headphone equalization, tracking latency, spatial resolution. The individual perceptual relevance of each element and appropriate error metrics are not clear.

The aim of the auralization system presented in this paper is to reproduce *in situ* a re-synthesized version of a loudspeaker in a room, over open headphones, with three degrees of rotational freedom. Such a system provides the possibility to seamlessly alternate between the real source present in the room, and the same source presented virtually over tracked headphones. We can then arbitrarily modify each of the elements of the audio rendering pipeline in order to study the effect of each component individually.

2. RE-SYNTHESIS

2.1 RIR Measurements

A single point-to-point multichannel room impulse response (MRIR) is required to re-synthesize a binaural room impulse response at a fixed location in a room. To this end, a swept sine signal is reproduced by a loudspeaker and recorded by an open microphone array. The room impulse response is obtained by deconvolving the original swept sine from the recording. The microphone array is composed of 6 miniature microphones (DPA 4060) arranged in pairs along perpendicular axes at 5 cm from an Earthworks M30 microphone at the center of the array. The holder for the microphones is 3D printed (see Fig. 1).

For validation and equalization, reference measurements are obtained at the receiver position using a binaural mannequin (G.R.A.S. KEMAR).

2.2 Spatial analysis

The analysis of the MRIR is based on the Spatial Decomposition Method (SDM) [1], and is implemented using the SDM Toolbox [7]. The concept behind SDM is based on the assumption that the measured soundfield can be represented as a succession of discrete acoustic events, each of them represented by a sample in the RIR with an associated direction-of-arrival (DOA). This assumption is certainly violated after more than one acoustic event arrive simultaneously at the receiver location, due to increasing

doi:10.25836/sasp.2019.31

Figure 1. Custom microphone array (5 cm radius).

echo density. However, increasing echo density eventually results in a diffuse sound field, which can be regarded as a succession of events with random DOA. The method has been successfully used to implement analysis and auralization of sound fields in a variety of scenarios, including concert halls [2], stage acoustics [3], or car cabin acoustics [4, 5], among others. In the present study, we use the broadband RIRs from the 7 microphones as input for the direction-of-arrival (DOA) estimation. The SDM analysis window is set to the smallest allowed size, and the resulting DOA vectors are smoothed using a moving average window of 16 samples (with a sampling rate of 48 kHz).

2.3 BRIR Synthesis

The SDM sound field parametrization consists of a [1×N] vector $\mathbf{p} = [p_1, p_2, ..., p_N]$ containing the pressure RIR and a [3×N] matrix $\mathbf{r} = [\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N]$ indicating the DOA for each of the samples in cartesian coordinates. A binaural reconstruction can be implemented in a straightforward manner by combining the DOA data and the pressure RIR with a Head-Related Impulse Response (HRIR) dataset. A binaural room impulse response (BRIR) can be reconstructed as a weighted summation of delayed HRIRs corresponding to the closest directions to each sound event. A set of BRIRs corresponding to arbitrary head rotations, can be created by rotating the DOA matrix.

At first, the indices $\hat{k}_n^u$ to the relevant HRIRs for each sound event in each head orientation, u, are found. This is done by finding the nearest HRIR for each sample, n, in the rotated DOA matrices $\mathbf{r}^u$.

$$\hat{k}_n^u = \underset{k \in 1,...,K}{\arg\min} \{ \mathrm{d}(r_n^u, \hat{\mathbf{r}}) \} \tag{1}$$

where $\hat{\mathbf{r}}$ is a $[3 \times K]$ matrix containing the source/receiver relative orientations of the HRIR dataset in cartesian coordinates and $\mathrm{d}(\cdot,\cdot)$ is the Euclidean distance. The rotated DOA matrices are created by multiplying $\mathbf{r}$ with corresponding rotation matrices.

$$\mathbf{r}^u = R_z(-\theta^u)\, R_y(-\phi^u)\, \mathbf{r} \tag{2}$$

where R_y and R_z rotation matrices to render an arbitrary head orientation (θ^u, ϕ^u). This applies to a right-hand rule coordinate system where positive Y is left and positive Z is up [1] .

[1] To achieve a correct head rotation the DOA has to be rotated in a reversed order. Roll rotation is excluded from the equation as it is not implemented in the framework.

Next, the BRIRs for all head orientations can be constructed using the indices $\hat{k}_n^u$ of the HRIRs, by delaying the HRIR at the nth positions by n samples and multiplying it by the instantaneous pressure, p_n:

$$\mathbf{BRIR}^u(t) = \sum_{n=1}^{N} p_n\, \mathbf{HRIR}_{\hat{k}_n^u} \circledast \delta(t-n), \tag{3}$$

where **HRIR** is a three-dimensional $[H \times K \times 2]$ matrix containing a HRIR dataset of H samples (per channel) and K source/receiver relative orientations, and t indicates the samples in the BRIR.

2.4 Reverb correction

A known artifact of SDM based auralizations is the increase of high frequency energy in the late reverb, leading to a whitening of the spectrum. A description of this problem and a time-frequency equalization approach was presented in [4]. This technique is especially useful for loudspeaker based auralizations and binaural renderings based on a virtual loudspeaker approach, as it uses the pressure RIR $\mathbf{p}$ as a reference. However, if a HRIR dataset with high spatial resolution is used - hence, the virtual loudspeaker set-up contains many virtual loudspeakers - the time-frequency filtering becomes time and resource consuming, as each separate stream needs to be equalized. For this reason, an alternative approach to compensate for the late reverberation coloration is presented below.

The main advantage of this approach is that only one reverberation equalization needs to be performed per head orientation, as it is performed on the re-synthesized BRIR. Two main assumptions are made here: 1) the reverberation time (RT_{60}) of all the re-synthesized BRIRs must be the same as the one from the pressure RIR, regardless of the head orientation and 2) the time-frequency deviations of the reverberation in the synthesized BRIRs are not time dependent. It is expected that these assumptions hold true in most situations and could be only violated in extreme cases of highly directional late reverberation or double decay slopes.

The first step is to decompose the pressure RIR, $\mathbf{p}$, and the re-synthesized BRIR, $\mathbf{BRIR}^u$ in fractional octave bands by using perfect reconstruction filters [6]. The implementation of the filters is largely based on that present in the SDM Toolbox [7]. Then, the frequency dependent RT_{60} of the pressure RIR, $\mathbf{RT}_{60,\mathrm{orig}}$ and the RT_{60} of the re-synthesized BRIR, $\mathbf{RT}_{60,\mathrm{resynth}}$ are computed. The computed frequency dependent RT_{60} is used to generate the parameters of an exponential function that is multiplied with each band of the re-synthesized BRIR to modify their reverberation time, following the method first presented in [8]. Finally, all of the corrected subbands $\mathbf{BRIR}^u_{\mathrm{corr},f}$ are summed together resulting in the corrected BRIR, $\mathbf{BRIR}^u_{\mathrm{corr}}$.

$$\mathbf{BRIR}^u_{\mathrm{corr}}(t) = \sum_{f=1}^{F} \mathbf{BRIR}_{\mathrm{corr},f}(t) \tag{4}$$

$$\mathbf{BRIR}^u_{\mathrm{corr},f}(t) = \mathbf{BRIR}^u_f(t) e^{-t(d_{1,f} - d_{0,f})} \tag{5}$$

 doi:10.25836/sasp.2019.31

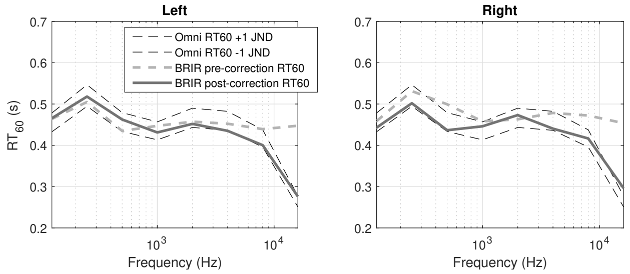


Figure 2. Comparison of the original RT_{60} and the re-synthesis before and after correction.

The constants of the exponential function, $d_{1,f}$ and $d_{0,f}$, are derived from the RT_{60} as follows

$$d_{0,f} = \frac{\ln(10^6)}{2\,RT_{60,\mathrm{resynth},f}} \tag{6}$$

$$d_{1,f} = \frac{\ln(10^6)}{2\,RT_{60,\mathrm{orig},f}} \tag{7}$$

A comparison between the estimated reverberation time of the original pressure RIR and the re-synthesized BRIRs is presented in Fig. 2. It is seen that before the correction the reverberation time is overestimated at high frequencies. After correction, the reverberation time falls within ± 1 JND (5% of the RT60, as defined in the ISO 3382).

2.5 Monaural equalization

The re-synthesized BRIRs can be regarded as a weighted summation of delayed HRTFs. Ideally, the spectral properties of the HRTF used for the re-synthesis should match perfectly with those of the subject listening to the re-synthesized BRIRs. However, the acquisition method of the HRTFs does in many cases introduce spectral distortions that are not easily characterized. For instance, in the case of simulated HRTFs, the skin absorption properties must be modeled, or in certain HRTF measurement systems the low frequency response must be extrapolated. For this reason, we use a mannequin measurement conducted at the listening position to generate a direction independent monaural filter that is applied to all the BRIRs at the reproduction stage. This equalization accounts as well for frequency response deviations of the omnidirectional RIR.

The filter is derived by dividing the smoothed magnitude response (1/12 octave band) of the left and right channels of a measured BRIR and its re-synthesis, ensuring that the relative source/receiver orientation is the same. The magnitude differences are averaged over left and right channels and a minimum phase FIR filter of 2048 taps is generated. Design of the filter is done using the `fdesign` and `rceps` functions of Matlab.

During experimental validation we found that the differences between equalization filters derived from various relative source/receiver orientations are negligible below approximately 12 kHz. At high frequencies, HRTF datasets are prone to present artifacts due to simulation limitations or spurious reflections in measurements. Thus we can consider this filter to be direction independent and the equalization is applied for any arbitrary orientation. Fig. 3 shows the monaural equalization filter response for various head orientations. In this case, the HRTF used for re-synthesis was generated from simulations, and thus a generalized spectral mismatch was expected.

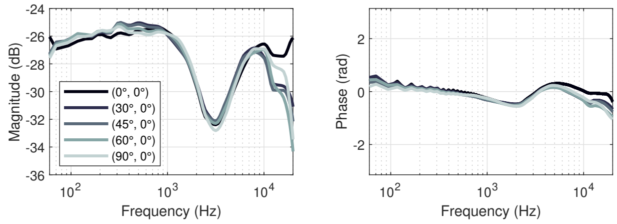


Figure 3. Magnitude and phase responses of a monaural equalization filter computed for various head orientations.

3. OBJECTIVE VALIDATION

An objective validation comparing measured BRIRs and their re-synthesized counterparts is presented in Fig. 4. Although the re-synthesis method does not aim at a physical reconstruction of the BRIRs, a close agreement is found between the original and the re-synthesized BRIRs. The results are presented for two rooms with different source/receiver orientations (frontal orientation in room A, and lateral orientation in room B).

The re-synthesized broadband pressure BRIRs resembles the overall time-energy structure of the original BRIRs, correctly displaying appropriate timing and amplitude of the most energetic events. However, some spurious reflections can be observed in the re-synthesis, most likely due to constructive interference of acoustic events when combining HRIRs for various directions.

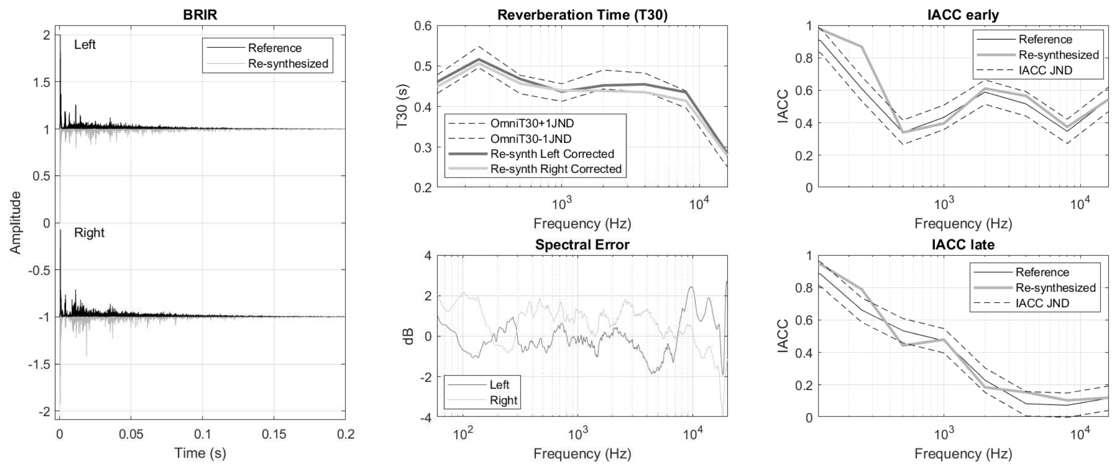
The RT_{60} of the re-synthesis fits within ± 1 JND (5% of the reference RT_{60}, as defined in ISO 3382) at most of the frequency bands.

The spectral error after monaural equalization falls within ± 2 dB at all frequencies up to 10 kHz. More variability is encountered between 10 and 20 kHz, likely due to uncertainties in the used HRTFs. The HRTF dataset used in this validation is simulated using the boundary element method (BEM) and a 3D mesh corresponding to the same mannequin used in the BRIR measurements (G.R.A.S. KEMAR).

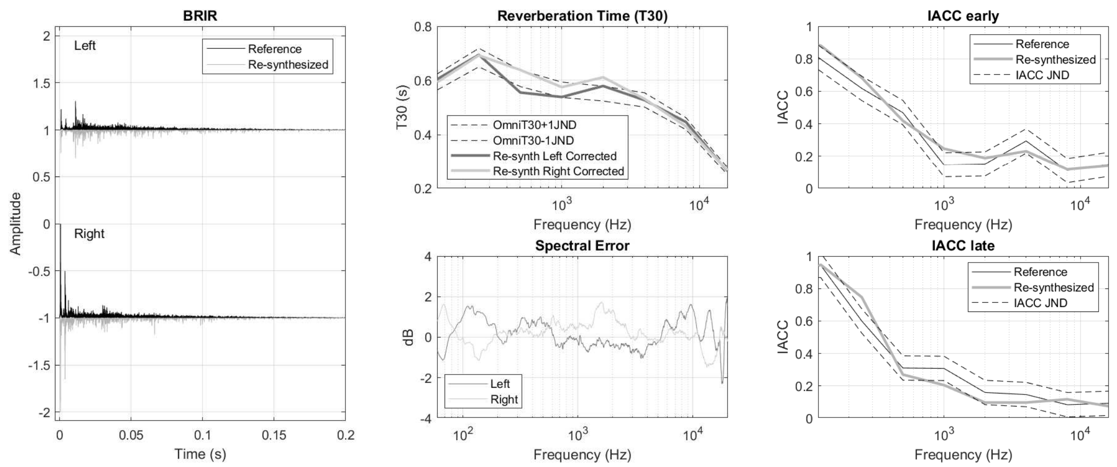
The early and late interaural cross correlation are generally close to ± 1 JND (0.075, as defined in ISO 3382) at all frequency bands, suggesting that the spatial properties of the re-synthesized BRIRs largely resemble those of the original BRIRs. Note that the late reverberation tails used in the re-synthesized BRIRs correspond to an arbitrary direction, different than those of the early reflections. This supports the assumption that the late reverberation can effectively be rendered as a direction independent filter [9].

4. REAL-TIME FRAMEWORK

An experimental framework has been developed in Max/MSP to handle the convolution of the BRIRs with an



a) Room A (frontal position).



b) Room B (lateral position).

Figure 4. Instrumental validation of the re-synthesized BRIRs as compared to the in-situ mannequin measurements.

anechoic signal, headphone equalization, monaural HRTF frontal/diffuse equalization and listener tracking. The framework can be interfaced with other Max/MSP patches allowing quick prototyping of listening tests by switching BRIR datasets and equalization filters seamlessly. The (end-to-end) motion to sound latency of the system is approximately 70 ms.

4.1 Convolution

The real-time operations for dynamic rendering are performed in Max/MSP. The convolution pipeline is split in 3 separate parts: direct sound, early reflections, and late reverberation. The direct sound (first 128 samples) and the early reflections are convolved dynamically, and filters are switched as head movements are detected. An equal power cross-fade with a duration of 1 ms is applied when switching filters to avoid audible artifacts. The convolution operations are performed using the spat5.conv˜ object.

If the mixing time is sufficiently high, the late reverberation can be modeled as a direction independent filter [9]. The framework implements a single binaural FIR convolution and the mixing time between early reflections and late reverberation can be arbitrarily adjusted. A window of variable size is implemented to ensure a smooth transition between the early and late part of the BRIR.

4.2 Tracking

The framework handles real-time tracking of the listener orientation and loudspeaker positions, ensuring that the relative angles between the listener and the real and virtual sources match. The loudspeaker tracking is implemented using an optical tracking system (OptiTrack or Vicon), while listener tracking is done either with an optical system or an Oculus Rift headset.

5. PERCEPTUAL VALIDATION

A pilot study to assess the authenticity of the re-synthesized BRIRs has been conducted, and a second test assessing plausibility is currently in progress. The authenticity test was composed of three parts: a discrimination test (ABX), an identification test (2 alternative forced choice - 2AFC) and a qualitative rating. Ten expert listeners wearing non-occluding headphones (AKG K1000) were presented with sounds coming from a loudspeaker,

 doi:10.25836/sasp.2019.31

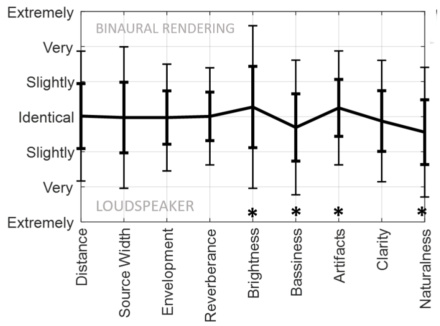


Figure 5. Perceptual ratings of the binaural renderings as compared to a real loudspeaker.

and a re-synthesized version using individualized HRTFs and headphone equalization. They were allowed to switch between presentation modes in real time and were granted unlimited listening time. In addition, the loudspeakers were visible. Three source positions and four sound samples were used (male and female speech, solo guitar, and pink noise), resulting in 12 trials. In these conditions, listeners were able to correctly discriminate in 114/120 (95%) of the trials, and to correctly identify which sample was the loudspeaker in 96/120 (80%) of the trials. However, only ratings of timbral cues, *naturalness* and *artifacts* presented statistically significant differences between the real loudspeaker and the renderings. In addition, several participants mentioned that in some cases a subtle localization mismatch between real and virtual sources was enough to inform their judgment in identifying which sample was the reference, and whether it was a real loudspeaker or a binaural rendering. A summary of the test results is presented in Fig. 5. It is worth noting that the participants in this study were highly familiar with binaural audio rendering and critical listening and lower discrimination and identification rates could be expected with naive listeners. To our knowledge, only one study assessed authenticity of individualized dynamic binaural audio, and although the BRIRs were generated using binaural microphones in the subjects' ears, authenticity was not achieved [10].

A pilot listening test was conducted to assess the plausibility of generic binaural renderings. In this case, a loudspeaker was hidden behind an acoustically transparent curtain and compared against a binaural re-synthesis generated with BEM simulated KEMAR HRTFs and generic headphone equalization (performed on the same mannequin). Seven expert listeners were granted unlimited time to listen to the two stimuli (castanets of approximately 7 seconds of duration) and were able to switch between them in real time. Then, they were asked which of the two sounds was perceived as more plausible - or likely to be a loudspeaker. Preliminary results suggest that in the absence of visual cues, and even with generic HRTFs, the perceived plausibility of virtual sounds is as high as that perceived for a real loudspeaker. At the moment of writing this manuscript a full test is being conducted to confirm these initial results.

6. BRIR MANIPULATIONS

The premise behind the implementation of this auralization system is to enable perceptual research in room acoustics. This section presents a variety of manipulations that we implemented in order to test the perceptual effects of various parts of the rendering pipeline.

6.1 Pre-rendered manipulations

A number of straightforward modifications in the re-synthesis algorithm allow the rendering of time-energy and spatial manipulations on the generated BRIRs. Those manipulations that require the modification of the HRTF dataset, the pressure RIR or the DOA vectors must generally be performed offline, prior to the rendering of re-synthesized BRIRs.

6.1.1 HRTF Dataset

The re-synthesis minimizes the distance between the DOA of an acoustic event and the HRIR used to render such event. Arbitrary manipulations of the spatial grid of the HRTF can be used to investigate the requirements of spatial resolution of HRTFs. Minimizing the size of a HRTF dataset to decrease memory requirements is important in mobile systems rendering spatial audio in real time.

6.1.2 DOA Quantization

The DOA vectors generated by SDM can be directly manipulated to render spatial manipulations of the sound field. For instance, quantizing the DOA vectors in order to limit the possible DOA of reflections while keeping intact the DOA of the direct sound can be useful to investigate the spatial resolution needed for the reproduction of reflections. In real-time systems, minimizing the number of reflection directions can contribute to decreasing the number of convolutions, optimizing compute requirements.

6.1.3 Reverberation time

The strategy introduced in [8] and used in Section 2.4 to correct the RT_{60} of re-synthesized BRIRs can be applied to implement arbitrary manipulations of the reverberation time. For example, modifying the reverberation time of a rendered BRIR and performing in-situ comparisons against a real loudspeaker allows the study of perceptual thresholds of the room acoustic divergence effect [11].

6.1.4 Synthetic reverberation

The late reverberation tail is rendered as a direction independent filter and it can be easily swapped with a synthetic reverberation. For instance, it is common for real-time sound engines to use reverberators for the synthesis of late reverberation. By combining spatial data with a synthetic monaural reverberation tail, various architectures can be compared.

6.2 Real-time manipulations

Those manipulations that are related to the real-time convolution of the BRIRs with anechoic audio can be rendered in real-time, as they not require the modification of any data prior to the re-synthesis of the BRIRs.

6.2.1 Relative levels of the RIR

Another manipulation related to the study of the room divergence effect [11] is that of the relative levels between the direct sound, early reflections and late reverberation. As the re-synthesized BRIRs are convolved in three separate pipelines, it is straightforward to modify the relative levels of each part of the BRIR. This is especially useful to study the link between the direct-to-reverberant ratio (DRR) and the perception of distance.

6.2.2 Direction dependent reverberation level

In [12], the importance of reverberation at the ipsi- and contralateral ear and their role in perceived externalization was investigated. However, the study was limited to static sources. The rendering approach presented here allows for the implementation of dynamic and direction dependent modifications of the reverberation level at each ear.

6.2.3 Distortion of the acoustic space

In anechoic conditions, the perceived angle of a sound source suffers angle-dependent distortions [13], i.e., the perceived direction does not correspond to the actual direction of reproduction. A module to compensate for spatial distortions is included to enable further investigation in reverberant conditions.

7. SUMMARY AND CONCLUSIONS

A framework for the re-synthesis of BRIRs based on SDM analysis is presented. An alternative approach to correct the known reverberation artifacts of SDM auralizations is introduced and validated. The auralizations can be dynamically reproduced using head tracking in a Max/MSP framework. An objective validation of the re-synthesized BRIRs by directly comparing them to measured BRIRs suggests that the time-energy, spatial and spectral properties are largely replicated, with minor deviations. The renders are found to be perceptually plausible when compared to real loudspeakers, although not indistinguishable from an explicit reference. Finally, a variety of manipulations of the re-synthesized BRIRs are introduced, enabling the study of BRIR degradations in direct comparison to real loudspeakers. Future work will focus on the study of perceptual thresholds in the degradation of BRIRs in order to establish the extent of acceptable acoustic deviations in the rendering of virtual sounds in augmented reality.

8. REFERENCES

[1] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, “Spatial decomposition method for room impulse responses,” *J. Audio Eng. Soc*, vol. 61, no. 1/2, pp. 17–28, 2013.

[2] J. Pätynen, S. Tervo, and T. Lokki, “Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses,” *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 842–857, 2013.

[3] S. Amengual Garí, M. Kob, and T. Lokki, “Investigations on stage acoustic preferences of solo trumpet players using virtual acoustics,” Proceedings of the 14th Sound and Music Computing Conference 2017, pp. 8; 167–174, 2017-07.

[4] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, and T. Lokki, “Spatial analysis and synthesis of car audio system and car cabin acoustics with a compact microphone array,” *J. Audio Eng. Soc*, vol. 63, no. 11, pp. 914–925, 2015.

[5] N. Kaplanis, S. Bech, S. Tervo, J. Pätynen, T. Lokki, T. van Waterschoot, and S. H. Jensen, “Perceptual aspects of reproduced sound in car cabin acoustics,” *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 1459–1469, 2017.

[6] J. Antoni, “Orthogonal-like fractional-octave-band filters,” *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 884–895, 2010.

[7] S. Tervo, “SDM Toolbox.” `https://www.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox`. Accessed: 2019-06-20.

[8] D. Cabrera, D. Lee, M. Yadav, and W. L. Martens, “Decay envelope manipulation of room impulse responses: Techniques for auralization and sonification,” in *Proceedings of Acoustics 2011*, (Gold Coast, Australia), 2011.

[9] A. Lindau, L. Kosanke, and S. Weinzierl, “Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses,” *J. Audio Eng. Soc*, vol. 60, no. 11, pp. 887–898, 2012.

[10] F. Brinkmann, A. Lindau, and S. Weinzierl, “On the authenticity of individual dynamic binaural synthesis,” *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1784–1795, 2017.

[11] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, “A summary on acoustic room divergence and its effect on externalization of auditory events,” in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, June 2016.

[12] S. Li, R. Schlieper, and J. Peissig, “The effect of variation of reverberation parameters in contralateral versus ipsilateral ear signals on perceived externalization of a lateral sound source in a listening room,” *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 966–980, 2018.

[13] W. O. Brimijoin, “Angle-dependent distortions in the perceptual topology of acoustic space,” *Trends in Hearing*, vol. 22, pp. 1–11, 2018.

doi:10.25836/sasp.2019.31