A System for Binaural Reproduction of Self-Generated Sound in VAEs

Johannes M. Arend[1,2], Philipp Stade[1,2], Christoph Pörschmann[1]
[1] *TH Köln, Institute of Communications Engineering, Cologne, Germany*
[2] *TU Berlin, Audio Communication Group, Berlin, Germany*
Email: johannes.arend@th-koeln.de

Introduction

Immersion is an important aspect of virtual acoustics. However, in most virtual acoustic environments (VAEs), the user is only a passive listener in a predefined scene and, depending on the features of the system, has the possibility to move within this scene or to change several acoustic properties of the room. This implies that usually acoustic interaction between the user and the virtual room is very limited or not possible at all. In particular, most systems do not allow any interaction with the virtual room by means of self-generated sound, like the own voice for example, even though there is evidence that adequate reproduction of self-generated sound affects the user's perception and might even enhance immersion and presence [1][2].
In this paper, we present a VAE system that is able to capture and reproduce self-generated sound in real time. Hence, the VAE is supplemented with a reactive component that feeds self-generated sound back into the virtual room and provides the acoustic response to the actions of the user. With the term self-generated sound, we refer to any self-generated organic signal (e.g. speech, singing, oral sounds or hand claps) as well as to any user-generated sound, for example playing an instrument. Thus, technically speaking, the system presented here generally works with any arbitrary sound source. Moreover, it considers the varying directivity of the user or the sound source in real time. These are two major differences compared to the few reactive VAEs introduced so far, which always assume that the sound source has a constant directivity and only cover a specific use case, like the reproduction of one's own voice [3][4][5] or of a certain musical instrument [6].
This paper is structured as follows. The first section describes the basic idea of the reactive VAE. Next, the methods section explains the implementation of the system with the corresponding processing steps. Finally, the conclusion gives a brief overview of the system and outlines possible applications, recent work, and future research questions.

Basic Idea

The purpose of the system is a room-related and plausible reproduction of self-generated sound in a headphone-based VAE. Figure 1 illustrates the functional schematic of the presented system. Similar to this schematic, the following section outlines the approach starting from the acting subject. In the first place, the user, which means the acting subject, generates an arbitrary sound.

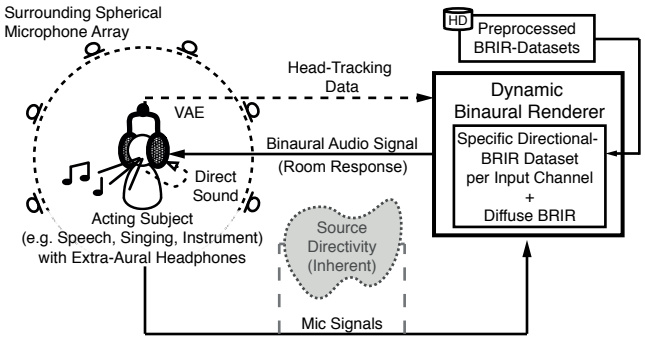

Figure 1: Functional schematic of the reactive VAE.

A spherical microphone array surrounding the user captures this direction-dependent sound in real time. The resulting microphone signals, which naturally contain the frequency-dependent directivity of the sound, now go to another essential part of the system: the dynamic binaural renderer. This real-time renderer convolves each microphone signal (or input channel) with a specific binaural room impulse response (BRIR). The respective *directional BRIRs* are preprocessed impulse responses describing a room-related direction-dependent binaural reflectogram per channel (see the methods sections for a more detailed explanation). Since the system applies dynamic binaural synthesis, the renderer requires an appropriate dataset of directional BRIRs per input channel. In addition, the renderer simultaneously convolves the sum of all microphone signals with a preprocessed *diffuse BRIR*, providing the reverberation of the simulated room. In order to maintain the natural direct sound of the user, the resulting binaural signal, which is composed of a direction-dependent reflection part (without direct sound) and a diffuse reverberation part, is played to the user over extra-aural headphones. By the use of such headphones, the direct sound can reach the ear of the user more or less unaffected, where it finally merges with the corresponding artificial room response. As usual in dynamic binaural synthesis, the renderer generates the binaural room response depending on the head orientation of the user, which is provided by a head tracker.

Methods

The presented reactive VAE combines different newly developed hardware and software components with available standard components. The following section outlines the most essential parts of the system and describes the methods of implementation.

Microphone Array The surrounding spherical microphone array serves as a tool to capture the user-generated sound with its specific directivity. The array has 32 channels and a diameter of 2 m. Its basic design is inspired by the spherical array from Pollow et al. [7]; a design which has been proven to be feasible for directivity measurements. The original shape of the array is a pentakis dodecahedron with 32 vertices, 60 faces and 90 edges. It is constructed out of fiberglass rods ($\varnothing = 6$ mm) and appropriately angled ABS plastic connectors. The connectors, which represent the vertices of the original shape, also serve as microphone holders for the 32 microphones (Rode NT5-S). Each connector is covered with a foam absorber to reduce interference artifacts. The whole structure stands on 20 cm stilts, which are also covered with foam; thus, the center of the array is at a height of about 1.20 m. For additional stability, the structure is tied with an aluminum truss system. Figure 2 shows a picture of the entire construction. As can be seen, we chose the dimensions of the array also with regard to the anechoic chamber, which has a relatively low ceiling height of about 2.30 m.

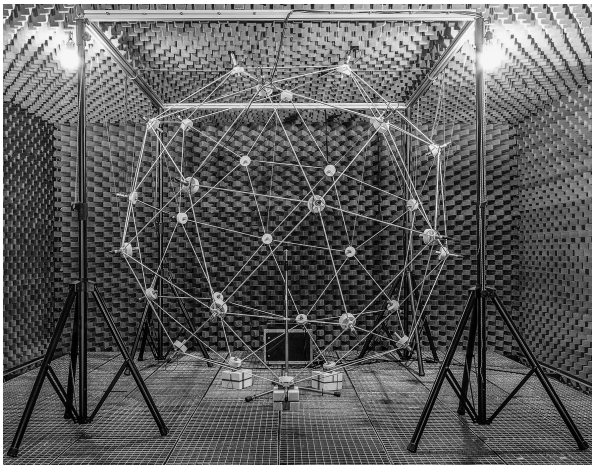

Figure 2: 32-channel surrounding spherical microphone array in the anechoic chamber at TH Köln.

Directional-BRIR Synthesis Besides the array, the specific directional-BRIR dataset per microphone channel is another key part of the system. Each fully synthesized directional BRIR basically describes a direction-dependent room response. The synthesis is based on room acoustic simulations with RAVEN [8]. Thus, as a start, the room must be modeled in 3D. Additionally, if acoustic measurements of the real room are available, the reverberation time of the model is fitted to the real reverberation time in an iterative process. Now, the room response (without direct sound) is simulated with a combination of the image-source method and ray tracing. In the simulation, the omnidirectional sound source and the receiver are placed at (almost) the same position. For further processing, which we implemented in Matlab, the list of results only from the image-source simulation is considered. This list specifies the delay, the outgoing angle from the source, the angle of incidence at the receiver and the frequency-dependent damping factors (in 1/3 octave bands) of each audible image source. Now, the basic principle is to assign every outgoing sound ray, which later leads to an incident reflection, to a predefined segment of the microphone array. Such a segment corresponds to the surface element (or face) allocated to a microphone (see Figure 3 (a)). In case of the pentakis dodecahedron, it is relatively easy to determine these faces by means of its dual polyhedron, which is the truncated icosahedron. In that regard, the 32 microphones are placed at the center of the 20 hexagon and 12 pentagon faces. Notionally, we now surround the outgoing rays with the segmented sphere, with source and receiver placed at the center of the sphere (see Figure 3 (b)). In this processing step, the algorithm assigns every outgoing ray to a segment through intersection point calculation (see Figure 3 (c)). This leads to a list of the related incident reflection rays per segment (see Figure 3 (d)). Thus, to each microphone, these reflections are assigned which would occur when the room would be excited only through the respective segment with an ideal loudspeaker, directed towards the center of the segment and with a directivity (or beam) according to the solid angle of the segment.

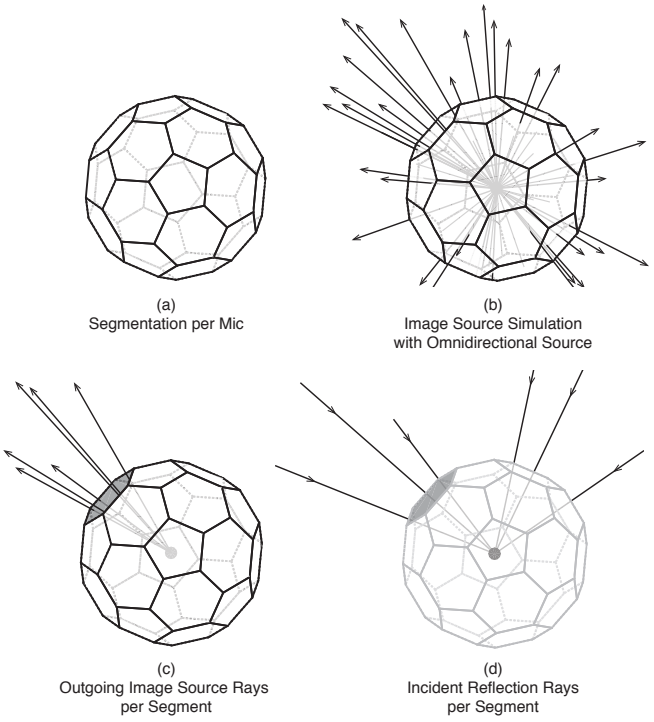

Figure 3: Illustration of the ray-segment assignment process.

After the successful assignment of the incident reflection rays to the segments, the actual synthesis of the directional BRIRs per microphone channel follows, which is a quite common process. According to the reflection list, the algorithm generates a synthetic directional BRIR by summing delayed, intensity-scaled, and filtered head-related impulse responses (HRIRs). The used HRIRs come from full spherical measurements on a Lebedev grid with 2702 nodes [9], which were transformed to the spherical harmonics domain and thus are stored as spherical harmonic coefficients. The required directions are then

extracted through spherical harmonic interpolation. The reflection filters are based on the frequency-dependent damping factors in 1/3 octave bands. These 31 values are first inter- and extrapolated to the desired number of filter taps and to the frequency range from 0 Hz up to half the sample rate. The reflection filters are then designed as Hann-windowed minimum and linear phase FIR filters, thus the filter type and the filter kernel size can be chosen appropriately. Moreover, latency compensation can be applied in this context, simply by subtracting the previously determined latency value from the delay value of each reflection.
The algorithm repeats this procedure for each required head orientation according to the used spatial grid. In our case, the binaural renderer works with a resolution of 1° in the horizontal plane, thus a directional-BRIR dataset per microphone contains 360 BRIRs. However, synthesizing directional-BRIR datasets for binaural rendering that involves vertical head movements is also possible. Figure 4 (top) again summarizes the described processing chain.

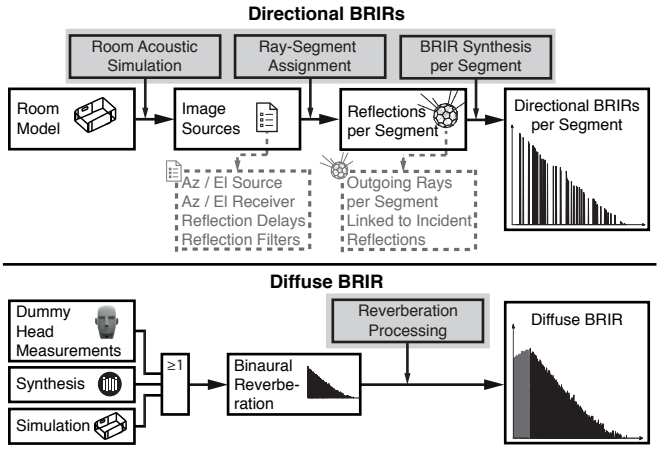

Figure 4: Processing chain for synthesizing the directional (top) and the diffuse (bottom) BRIRs.

Diffuse-BRIR Synthesis The reverberant part of the BRIRs is generally assumed to be diffuse and thus should not have any directional information. Therefore, only one appropriate binaural reverberation impulse response is applied. As Figure 4 (bottom) shows, the binaural reverberation can be measured with a dummy head in the real room, it can be fully synthesized [10], or it comes from the results of the ray-tracing simulation. Depending on the reverberation source, the processing in Matlab involves different steps. If a measured BRIR is used, the diffuse part after the (perceptual) mixing time is extracted and leveled appropriately with respect to the simulated reverberation. Because one of our recent studies showed that perceptual results of auralizations with synthetic BRIRs are significantly better when the early part of the BRIR also contains diffuse components [11], the section before the mixing time can optionally be filled with diffuse reverberation, which primarily works as a masker filling the gaps between the single reflections. Using fully synthetic reverberation is another option. In this case, the binaural reverberation is based on frequency-dependent shaped noise, which is leveled appropriately and which can be spread over the entire time range if desired. The simplest way is to use the results of the ray-tracing simulation, since no further processing or additional leveling has to be applied. In a final step, the respective binaural reverberation is scaled in intensity according to the number of feeding microphone channels.

Compensation Filters The system relies on specific filters to compensate for the magnitude response of both, the extra-aural headphones (AKG K1000) and the array microphones (Rode NT5-S). Concerning the headphones, we determined the compensation filters in the usual way [9] for two reproducible earphone positions (fully open and closed). For the microphone compensation filters, we measured four different array microphones and compared them to a reference measurement with an Earthworks M30 microphone. The averaged spectral differences between the array microphones and the reference microphone yielded the magnitude response of the inverse filter. Both of these final compensation filters are available as minimum and linear phase Hann-windowed FIR filters and can be applied to the final BRIR datasets.

Setup The final setup of the reactive VAE is relatively straightforward, as can be seen from the block diagram in Figure 5. The 32 Rode NT5-S microphones are connected to four RME Octamic II preamps and AD converters, which again are connected to two RME Fireface UFX audio interfaces. The two interfaces work together as one aggregate device in the iMac PC. With the JACK Audio Connection Kit, the 32 channels are fed to 32 corresponding directional-BRIR sources in the SoundScape Renderer [12]. Furthermore, the 32 channels are summed to another channel, which is fed to the additional diffuse-BRIR source in the renderer. A Polhemus Fastrak provides the head-tracking data so that the renderer can generate the binaural signal according to the actual head orientation of the user. The binaural signal is then DA converted with the main RME Fireface UFX interface, amplified with an Harman HK650 amplifier and finally played to the user over the AKG K1000 headphones.

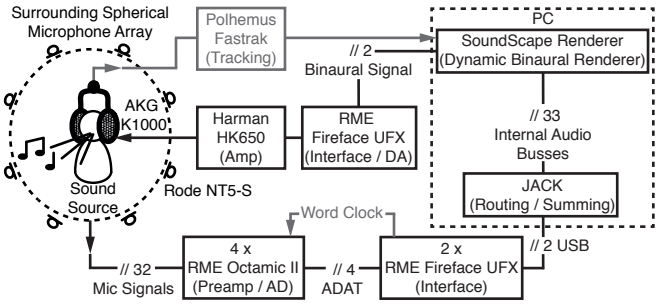

Figure 5: Setup of the reactive VAE.

Level Calibration Since a correct level ratio between direct sound and synthesized room response is crucial for an accurate reproduction, the entire system has to be calibrated. Of course, first of all, the input gain of the microphones as well as the level of the BRIRs must be set to reasonable values in order to achieve an overall good

signal-to-noise ratio. The general idea of the calibration method, which is related to the approach from Böhm et al. [6], is to adjust the playback level of the system so that a real acoustic scene in the anechoic chamber and a binaural simulation of this scene lead to the same RMS level. For this purpose, we placed a loudspeaker at the center of the array and the KU100 dummy head at a distance of 5 m. We than played a pink noise sequence over the loudspeaker and simultaneously recorded this with an array microphone located in front of the loudspeaker as well as with the dummy head. Next, we auralized this scene with the microphone recording as the source signal and a simulated (anechoic) BRIR as the spatial filter, and recorded the binaural signal (played through the headphones) with the dummy head. In the last step, we matched the RMS level of the recorded auralization with the RMS level of the real dummy head measurement. To reduce the influence of the anechoic chamber and the array construction, we filtered both signals with a 48 dB/octave band-pass filter ($f_l = 500\,\text{Hz}$, $f_h = 5\,\text{kHz}$). This allowed for a better comparison of the RMS levels. As a result of the calibration, the real scene and the synthesized counterpart provide the same RMS level, which again results in a correct level ratio between direct sound and synthesized room response, at least as long as the direct sound level in the simulation remains unchanged.

Conclusion

These days, most VAEs do not allow acoustic interaction between the user and the virtual room by means of any self-generated sound. The few systems introduced so far which are able to do so, however, cannot handle arbitrary sound sources with varying directivity and are mostly designed for only one specific use case. In this paper, we presented a headphone-based reactive VAE which is able to capture and reproduce any self-generated sound in real time. As described in the methods section, the most important parts of the system are the 32-channel surrounding spherical microphone array used to capture the user-generated sound with its specific directivity, and the synthesized BRIRs, which contain the direction-dependent reflections (directional BRIRs) as well as the binaural reverberation of the room (diffuse BRIR). Furthermore, we depicted the headphone and microphone compensation as well as the level calibration. The final setup, with the array and the dynamic binaural renderer as key components, is relatively straightforward and provides a plausible binaural reproduction of self-generated sound.
The reactive VAE can be used, for example, as a virtual practice room for musicians, as part of an interactive VR application, or as a research tool. In recent work, we have implemented a shoebox-shaped test room and a concert hall. A first technical evaluation yielded proper functioning of the system. Several informal listening experiments revealed that the reactive VAE performs well and produces perceptually pleasant results.
Our future work will first focus on a more detailed technical evaluation of the system. As part of this, we plan to investigate different scenarios where the sound source is spatially extended or not in the center of the array. Besides these technical issues, we are going to examine the influence of self-generated sound on human perceptual processes. This could be specific studies concerning the spatial resolution or the number of early reflections required for self-generated sound in comparison to external sound, or studies investigating the influence of self-generated sound on attributes like immersion and presence.

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (BMBF) under the support code 03FH014IX5-NarDasS. We thank Jonas Christofzik for his great help in building up the array.

References

[1] Pörschmann, C. and Pellegrini, R. S., "3-D Audio in Mobile Communication Devices: Effects of Self-Created and External Sounds on Presence in Auditory Virtual Environments," *JVRB - Journal of Virtual Reality and Broadcasting*, 7(2010)(11), pp. 3–11, 2010.

[2] Nordahl, R. and Nilsson, N. C., "The Sound of Being There: Presence and Interactive Audio in Immersive Virtual Reality," in K. Collins, B. Kapralos, and H. Tessler, editors, *The Oxford Handbook of Interactive Audio*, chapter 13, pp. 213–233, Oxford University Press, New York, USA, 2014.

[3] Pörschmann, C., "One's Own Voice in Auditory Virtual Environments," *Acta Acust united Ac*, 87(3), pp. 378–388, 2001.

[4] Yadav, M., Cabrera, D., and Martens, W. L., "A system for simulating room acoustical environments for one's own voice," *Applied Acoustics*, 73(4), pp. 409–414, 2012.

[5] Pelegrín-García, D., Rychtáriková, M., Glorieux, C., and Katz, B. F. G., "Interactive auralization of self-generated oral sounds in virtual acoustic environments for research in human echolocation," in *Proc. of Forum Acusticum*, pp. 1–6, 2014.

[6] Böhm, C., Schärer Kalkandjiev, Z., and Weinzierl, S., "Virtuelle Konzerträume als Versuchsumgebung für Musiker," in *Proc. of the 42nd DAGA*, pp. 833–835, 2016.

[7] Pollow, M., Behler, G., and Masiero, B., "Measuring Directivities of Natural Sound Sources With a Spherical Microphone Array," in *Proc. of the Ambisonics Symposium, Graz*, pp. 1–6, 2009.

[8] Schröder, D. and Vorländer, M., "RAVEN: A Real-Time Framework for the Auralization of Interactive Virtual Environments," in *Proc. of Forum Acusticum*, pp. 1541–1546, 2011.

[9] Bernschütz, B., "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," in *Proc. of the 39th DAGA*, pp. 592–595, 2013.

[10] Stade, P. and Arend, J. M., "Perceptual Evaluation of Synthetic Late Binaural Reverberation Based on a Parametric Model," in *Proc. of the AES International Conference on Headphone Technology, Aalborg, Denmark*, pp. 1–8, 2016.

[11] Stade, P., Arend, J. M., and Pörschmann, C., "Perceptual Evaluation of Synthetic Early Binaural Room Impulse Responses Based on a Parametric Model," in *Proc. of the 142nd AES Convention, Berlin, Germany*, pp. 1–10, 2017.

[12] Geier, M., Ahrens, J., and Spors, S., "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," in *Proc. of the 124th AES Convention, Amsterdam, The Netherlands*, pp. 1–6, 2008.