Binaural reproduction of self-generated sound in virtual acoustic environments

Johannes M. Arend, Philipp Stade, and Christoph Pörschmann

Citation: Proc. Mtgs. Acoust. **30**, 015007 (2017); doi: 10.1121/2.0000574
View online: http://dx.doi.org/10.1121/2.0000574
View Table of Contents: http://asa.scitation.org/toc/pma/30/1
Published by the Acoustical Society of America

Proceedings of Meetings on Acoustics

Volume 30 http://acousticalsociety.org/

Acoustics `17 Boston

173rd Meeting of Acoustical Society of America and 8th Forum Acusticum
Boston, Massachusetts
25-29 June 2017

Architectural Acoustics: Paper 4aAAd2

Binaural reproduction of self-generated sound in virtual acoustic environments

Johannes M. Arend and Philipp Stade
Institute of Communications Engineering, TH Köln, Köln, NRW, 50679, GERMANY; TU Berlin, Audio Communication Group, Berlin, GERMANY; Johannes.Arend@th-koeln.de; Philipp.Stade@th-koeln.de

Christoph Pörschmann
Institute of Communications Engineering, TH Köln, Köln, NRW, 50679, GERMANY; Christoph.Poerschmann@th-koeln.de

User interaction with the virtual acoustic environment (VAE) has become of increasing interest in research. However, so far little attention has been paid to interaction by means of self-generated sound, like, for example, the own voice, even though this would open up new possibilities for natural interaction. Moreover, there is evidence that adequate reproduction of self-generated sound might enhance presence. This paper presents a system that allows such interaction with the VAE. For this, the so-called reactive VAE captures the self-generated sound, feeds it back into the virtual room, and provides the acoustic response to the actions of the user in real time. The specific features are that the system considers the dynamic directivity of the sound source and that it generally works with any source. The paper describes the implementation of the system, a technical evaluation that confirms basic functionality, and two implementation examples. Moreover, the study discusses technical general conditions of the system architecture. The potential applications of the reactive VAE are diverse, including use as a virtual practice room for musicians or as a research tool. In future work, the system will provide a basis for experiments investigating the influence of self-generated sound on human perceptual processes.

Published by the Acoustical Society of America

 https://doi.org/10.1121/2.0000574

1. INTRODUCTION

Interaction between the user and the virtual acoustic environment (VAE) has become of increasing interest in virtual acoustic research. Thus, as an example, more and more systems allow the user to move within the virtual room or to change several acoustic properties in real time.[1,2] However, so far, little attention has been paid to interaction with the VAE by means of self-generated sound, such as the own voice, even though this would open up entirely new possibilities for natural interaction with the virtual room. Moreover, there is evidence that adequate reproduction of self-generated sound (e.g. voice or footsteps) affects the user's perception and might even enhance immersion and presence.[3,4] Hence, the reproduction of self-generated sound could be another important component in obtaining immersive VAEs.

In this paper, we present a so-called reactive VAE, which allows acoustic interaction between self-generated sound of the user and the virtual room. For this, the system captures the self-generated sound, feeds it back into the virtual room, and provides the acoustic response to the actions of the user in real time. Here the term self-generated sound means any self-generated organic signal (e.g. speech, singing, oral sounds, or hand claps) as well as any interactive sound, like playing an instrument for example. Thus, from a technical point of view, the presented system generally works with any sound source. Moreover, it considers the dynamic (or varying) directivity of the user or of 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[5–7] or of a certain musical instrument.[8,9] It is important to point out that we plan to use the introduced reactive VAE primarily as a research tool. Thus, with the aid of the system, we are going to conduct studies on the influence of self-generated sound on human perceptual processes.

The paper is structured as follows. Section 2 outlines the basic idea of the reactive VAE. Next, section 3 describes the design and implementation of the system from the hardware and software side. Section 4 addresses a first technical evaluation of the system, which is partly based on example implementations of a shoebox-shaped test room and of a concert hall. Finally, section 5 concludes the paper with a brief overview of the system, a short discussion of possible applications, a summary of the technical evaluation results, and some thoughts about future work and research questions.

2. BASIC IDEA

The aim of the system is to provide a room-related and adequate reproduction of self-generated sound in a headphone-based VAE. Consequently, the user, or in other words the acting subject, is central to the approach. This can be seen in Fig. 1, which illustrates the functional schematic of the reactive VAE. In accordance with this schematic, the following section outlines the approach starting from the acting subject.

First of all, the user generates an arbitrary sound. A spherical microphone array surrounding the user captures this direction-dependent sound in real time. The resulting microphone signals, which inherently provide the frequency-dependent directivity of the sound, now go straight to the binaural renderer as well as to the ReSource module (ReSource - Reverberation Source). The ReSource module is a stand-alone software component which processes all incoming microphone signals and provides an adaptively filtered mono signal at the output, further called the *reverberation source signal*, which then is used as the excitation signal for the diffuse reverberation synthesis (see section 3.D for a description of the ReSource algorithm). The subsequent dynamic binaural renderer convolves each raw microphone signal with a specific *directional BRIR* (BRIR - Binaural Room Impulse Response) and the reverberation source signal with a single *diffuse BRIR*. The respective directional BRIRs are preprocessed impulse responses describing a room-related direction-dependent binaural reflectogram per microphone channel (see section 3.B for a more detailed explanation). Since the system applies dynamic binaural synthesis, the renderer requires a dataset with an appropriate number of directional BRIRs per microphone channel. The diffuse BRIR on the other hand describes the

isotropic binaural reverberation of the simulated room (see section 3.C for more information). In a last step, the resulting binaural audio signal, which is composed of a direction-dependent reflection part (without direct sound) and of a diffuse reverberation part, is played to the user over extra-aural headphones. By the use of such headphones, the natural direct sound of the user can be maintained and reaches the ear of the user more or less unaffected, where it finally merges with the corresponding artificial room response. As is 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.

As a side note, please remember that the ReSource module outlined above (and included in Fig. 1) was not part of the system as described in our previous publication.[10] However, our technical and perceptual evaluations showed that it is a necessary component for proper reverberation synthesis.

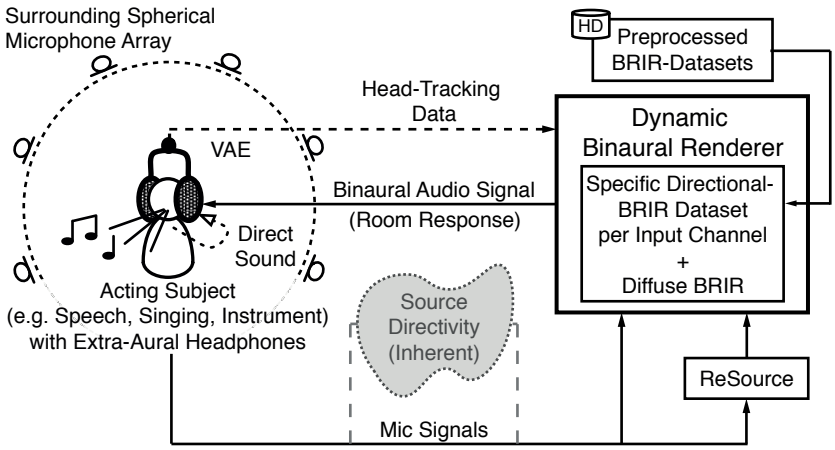

Figure 1: Functional schematic of the reactive VAE. The surrounding spherical microphone array captures the sound of the acting subject. The microphone signals go to the dynamic binaural renderer and to the ReSource module, which provides the reverberation source signal. The dynamic binaural renderer convolves each raw microphone signal with a specific directional BRIR and the reverberation source signal with a diffuse BRIR. The resulting binaural audio signal is played to the user over extra-aural headphones. According to 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.

3. DESIGN AND IMPLEMENTATION

The presented reactive VAE combines several newly developed hardware and software components with available standard components. The following first three subsections describe the design and implementation of the microphone array and of the BRIR datasets, which both clearly represent the centerpieces of the system. This is followed by a brief explanation of the ReSource module, which provides the proper excitation signal for the reverberation synthesis. The last three subsections outline the implementation of the headphone and microphone compensation filters, the final setup combining all components, and the procedure for the level calibration of the entire system.

A. MICROPHONE ARRAY

The 32-channel surrounding spherical microphone array allows to capture the direction-dependent sound radiated by the user. The basic design of the array is inspired by the construction from Pollow et al.,[11] which has been shown to be feasible for directivity measurements. The basic shape of the array is a pentakis

dodecahedron (32 vertices, 60 faces, 90 edges) with a diameter of 2 m. The structure is made of fiberglass rods ($\varnothing$ = 6 mm) and appropriately angled connectors, which were 3D-printed out of ABS plastic. The connectors represent the vertices of the original shape and serve as holders for the 32 Rode NT5 microphones. To reduce reflections and consequential interference artifacts, each connector is covered with a foam absorber. The whole structure stands on foam-covered stilts with a height of about 20 cm, leading to an array center 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, placed in the anechoic chamber at TH Köln. As can be seen, we chose the dimensions of the array also with regard to the size of the anechoic chamber, which has a relatively low ceiling height of about 2.30 m.

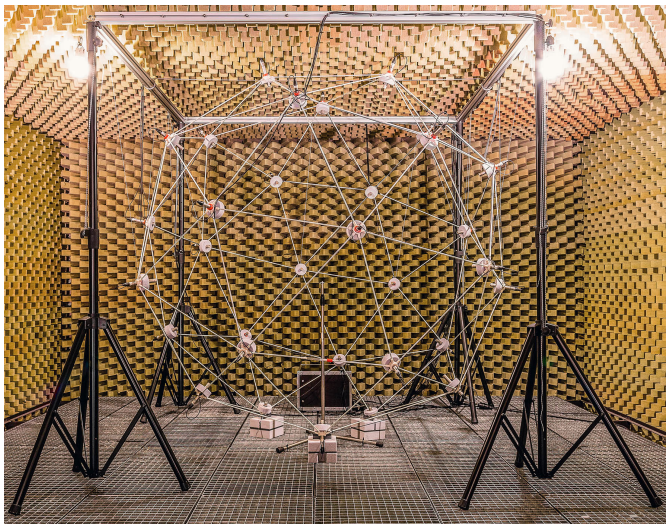

Figure 2: Surrounding spherical microphone array in the anechoic chamber at TH Köln. The array has 32 channels and a diameter of 2 m. Its basic shape is a pentakis dodecahedron.

B. DIRECTIONAL-BRIR SYNTHESIS

The specific directional-BRIR datasets are essential for an adequate reproduction of sound sources with dynamic directivity. Basically, each directional BRIR describes a direction-dependent room response by means of specular reflections. The synthesis is based on a room acoustic simulation with RAVEN,[12] which is why the room must be modeled in 3D first. Additionally, if acoustic measurements of the real room are available, the reverberation time of the model is matched to the measured reverberation time in an iterative process by fitting the absorption coefficients of the materials. Now, the room response (without direct sound) is simulated with a combination of the image-source method and ray tracing. Since the simulation is designated for self-generated sound, the omnidirectional sound source and the receiver are placed at (almost) the same position. All further processing including the actual synthesis is implemented in Matlab. For the directional-BRIR synthesis, only the list of results 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 each microphone (see Fig. 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 Fig. 3 (b)). In this processing step, the algorithm assigns every outgoing ray to a segment through intersection point calculation (see Fig. 3 (c)). This leads to a list of the related incident reflection rays per segment (see Fig. 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.

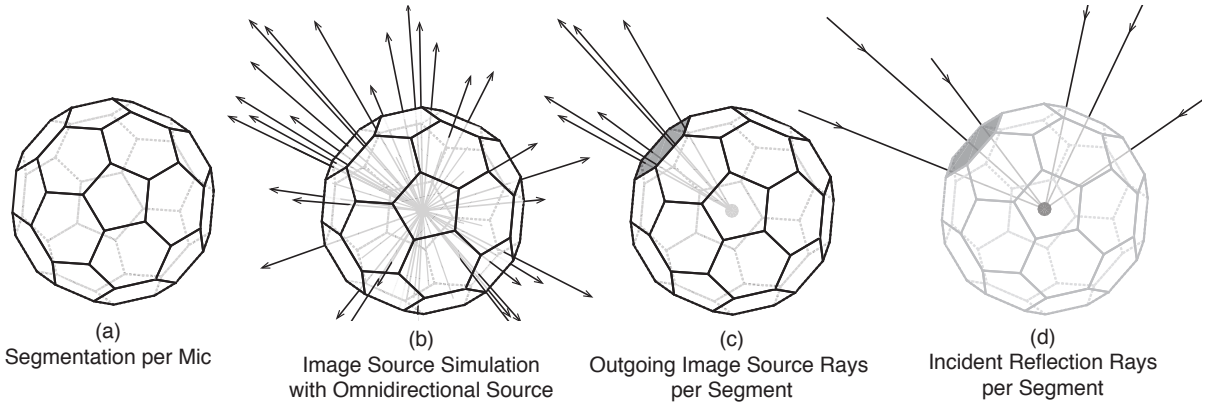

Figure 3: Illustration of the ray-segment assignment process. The basic principle is to assign every outgoing sound ray, which later leads to an incident reflection, to a predefined segment of the microphone array. The result is a list of the related incident reflection rays per segment.

Now that each outgoing sound ray (and the resulting incident reflection) is successfully assigned to one of the 32 segments, the actual directional-BRIR synthesis follows. In brief, the algorithm generates a synthetic directional BRIR by summing delayed, intensity-scaled, and filtered head-related impulse responses (HRIRs). The used HRIRs are based on spherical Neumann KU100 measurements made on a Lebedev grid with 2702 nodes.[13] Because arbitrary directions are needed for the reflection synthesis, the HRIR dataset is transformed to the spherical harmonics domain and thus stored as spherical harmonic coefficients. This way, the algorithm can extract any required direction through spherical harmonic interpolation.[14] Each HRIR is then filtered with a specific reflection filter, which basically describes the absorption properties of the simulated room. The filters are based on the frequency-dependent damping factors in 1/3 octave bands, obtained by the RAVEN simulation. 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. Each reflection filter is then designed as a Hann-windowed minimum and linear phase FIR filter, thus the filter type and the filter kernel size can be chosen appropriately. Moreover, audio-latency compensation can be applied in this context, simply by subtracting the previously determined audio-latency value from the delay value of each reflection. It is important to note that for a receiver and source height of about 1.20 m, the first reflection is usually the floor reflection with a delay of about 7 ms. However, depending on the buffer size, the audio latency is mostly higher than 7 ms. Therefore, the algorithm offers the option to skip the floor reflection in order to provide full audio-latency compensation, or to maintain the floor reflection and thus to compensate only the delay up to the first reflection.

The algorithm repeats the described 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. As a matter of course, any full spherical HRIR dataset can be used for the synthesis. Figure 4 (top) again summarizes the described processing chain.

Finally, it must be mentioned that we also considered measuring directional BRIRs instead of using the described simulation and synthesis approach. However, obtaining these BRIRs by means of acoustic measurements is hardly possible. First of all, an ideal sound source with steerable directivity would be necessary to excite the room at the different required directions. Moreover, since it is about self-generated sound, the source and the receiver (dummy head or microphone array) had to be placed at (almost) the same position, which causes further problems in implementation and measurement. Apart from that, with a dummy head or a sequential microphone array as the receiver, measuring impulse responses for all required excitation directions and receiver orientations would take a vast amount of time. For these reasons, we decided to use simulations with RAVEN to obtain the directional acoustic room properties.

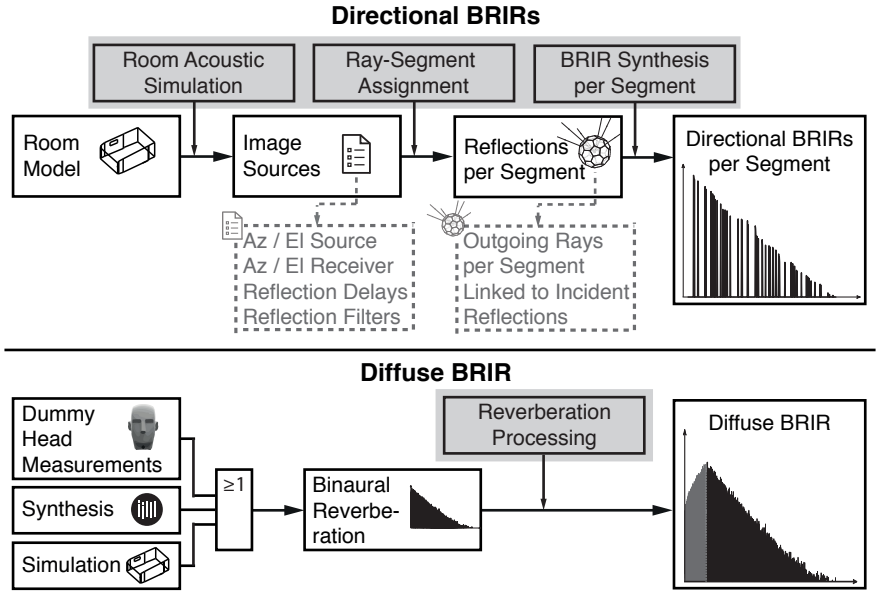

Figure 4: Processing chain for synthesizing the directional (top) and the diffuse (bottom) BRIRs. The directional BRIRs are based on a room acoustic simulation. The resulting image sources are assigned to the spherical segments, which leads to a list of reflections per segment. Finally, the directional BRIRs are synthesized by summing delayed, intensity-scaled, and filtered HRIRs. The single diffuse BRIR is based on dummy head measurements, synthesis, or simulation. Depending on the input, the reverberation processing involves different steps like energy matching, time shifting, and ramping, resulting in the final diffuse BRIR.

C. DIFFUSE-BRIR SYNTHESIS

As the term diffuse BRIR implies, we assume that the reverberant part of the virtual room is homogeneous and isotropic. Consequently, we apply only one appropriate binaural reverberation impulse response for the diffuse-BRIR synthesis. As Fig. 4 (bottom) shows, the binaural reverberation can be measured with a dummy head in the real room, it can be fully synthesized,[15] or it comes from the results of the ray-tracing simulation with RAVEN. Depending on the chosen reverberation source, the processing in Matlab involves different steps. Please note that on some points, the reverberation processing described in the following slightly differs from the processing introduced in our previous publication.[10] If a measured BRIR is used, the diffuse part after the (perceptual) mixing time is extracted and energy matched with respect to the simulated reverberation, which acts as the level reference. Next, the entire diffuse BRIR is shifted in time to the starting point of the first (latency-compensated) reflection. In the last step, a linear or cosine-squared ramp

is applied from the starting point up to the perceptual mixing time of the particular room, which is a procedure quite similar to the one introduced from Coleman et al.[16] In this way, the diffuseness slowly builds up through the early part and thereby works as a masker filling the gaps between the single reflections of the directional BRIR. This leads to significantly better perceptual results compared to conditions where the early part only contains specular reflections, as shown in one of our recent studies.[17] Using fully synthetic reverberation is another option. In this case, the binaural reverberation is based on frequency-dependent shaped noise, which can be spread over the entire time range if desired.[15] Generating the synthetic reverberation only requires the frequency-dependent reverberation time, which the simulation or measurements provide. Similar to the processing of the measured BRIR, the algorithm matches the energy with the simulated reverberation, shifts the entire BRIR to the starting point of the first reflection, and applies a linear or cosine-squared ramp up to the (perceptual) mixing time. The simplest way however is to use the results of the ray-tracing simulation, since no further processing or additional energy matching has to be applied.

D. RESOURCE MODULE

As briefly described in section 2, the reproduction system renders the directional and diffuse part of the binaural room response separately (see also Fig. 1). Regarding the directional part, the raw microphone signals can be passed straight to the binaural render for convolution with the respective directional BRIR. However, for the diffuse part, a specific *reverberation source signal* has to be acquired first. This source signal can be derived from the so-called *primal signal*, which contains all directional information of the source and therefore basically describes its full-spherical directivity. Zotter,[18] who introduced the term primal signal, discussed several approaches to obtain its spectrum and its much harder to determine phase information. In general, the primal signal is the superposition of all microphone signals. However, summing all microphone signals directly in time domain leads to interference artifacts because of slight time differences between the signals and thus is not feasible. One method to obtain the spectrum of the primal signal is to average the power spectra of all microphone signals. Indeed, this way all phase information of the primal signal gets lost, but for our specific use case, the phase information is not of importance. More precisely, we only need the spectrum of the primal signal in order to derive an appropriate excitation signal (the reverberation source signal) for the diffuse reverberation synthesis, as described in the following.

The basic idea behind the ReSource algorithm is to continually filter one microphone signal so that it matches the spectrum of the primal signal. This filtered signal is then passed to the binaural renderer as the reverberation source signal for convolution with the diffuse BRIR. Because both, the spectrum of the primal signal and of the microphone signal to be filtered permanently change for a rotating sound source with dynamic directivity, the filter has to be adapted continually. Moreover, all processing needs to be applied in real time. Thus, we implemented the algorithm as a stand-alone C++ module.

In brief, the processing is as follows. The 32 microphone signals at the input of the module are continually transformed to frequency domain with a short-time Fourier transform (STFT). For each frame of the STFT, the algorithm calculates the spectrum of the primal signal $S(n,k)$ by averaging the power spectra of all microphone signals, such as

$$S(n,k) = \sqrt{\sum_{i=1}^{N} w_i |(X_i(n,k)|^2} \tag{1}$$

where n is the time index, k is the frequency bin, i is the microphone channel number, N is the number of microphones (here N= 32), $X_i(n,k)$ denotes the STFT of the ith microphone signal, and w_i indicates the weight of the ith microphone channel, which correspond to the sphere surface area covered by the ith microphone (see segmentation per microphone in section 3.B). Next, the power spectra of the primal signal and of the microphone signal to be filtered are smoothed using logarithmic 1/3-octave spectrum smoothing.

The power spectrum of the adaptive filter $F(n,k)$ is then calculated such as

$$F(n,k) = \frac{S_{sm}(n,k)}{X_{i_{sm}}(n,k)} \tag{2}$$

where i is predefined with respect to one specific microphone whose filtered signal will be used as the reverberation source signal. Here, we use one of the microphones located at the top of the array. The reason is that, for a user rotating inside the array, signal changes in level and spectrum are relatively low at a top microphone compared to a microphone located in the horizontal plane for example. $F(n,k)$ now passes some regularization routines mainly for stabilization in the low and high frequency range. Next, the linear phase FIR filter $f(n)$ is obtained by inverse Fourier transform of the filter power spectrum $F(n)$ and by Hann windowing in time domain. Using the Hilbert transform, the filter is transformed to a minimum phase FIR filter. Finally, the algorithm convolves the signal from microphone i with the filter $f(n)$ by FFT-based convolution, providing the single-channel reverberation source signal at the output. According to this processing loop, the filter $f(n)$ automatically adapts to the changing spectra each STFT frame. Thus, the frame size is of particular importance because on the one hand, it must be large enough so that low frequency components can be detected, and on the other hand, it must be sufficiently short so that the filter can be adapted fast enough. However, as informal tests showed, the filter resulting for a rotating sound source inside the array is mostly some kind of smooth high-shelf filter changing relatively slow over time. Thus, the frame size can be quite large, which then leads to a moderate filter-adaption rate. Nevertheless, we need further informal listening tests for precise statements concerning frame size, hop size, and filter length.

E. COMPENSATION FILTERS

To compensate for the magnitude response of both, the extra-aural headphones (AKG K1000) and the array microphones (Rode NT5), we designed specific filters which can be directly applied to the BRIR datasets. Concerning the headphone compensation filters, we made 20 measurements with repositioning for two reproducible earphone positions each (fully open and closed) and used the algorithm from Bernschütz[14] for proper inversion. Regarding 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 measurements with the array microphones and the reference microphone yielded the magnitude response of the inverse filter. All compensation filters are available as minimum and linear phase Hann-windowed FIR filters.

F. SETUP

The final setup of the reactive VAE is relatively straightforward, as can be seen from the block diagram in Fig. 5. The 32 Rode NT5 microphones are connected to four RME Octamic II preamps and AD converters, which again are connected to two RME Fireface UFX audio interfaces. Both interfaces work together as one aggregate device in the iMac computer. All further internal routing is realized with the JACK Audio Connection Kit. For one thing, all 32 channels are passed through JACK directly to the corresponding directional-BRIR sources in the SoundScape Renderer.[19] For another thing, the 32 channels are routed to the ReSource module, which returns the single-channel reverberation source signal to JACK. This signal is then passed 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 head orientation of the user. Finally, the binaural signal is DA converted with the main RME Fireface UFX interface, amplified with a Harman HK650 amplifier, and played to the user over the AKG K1000 headphones.

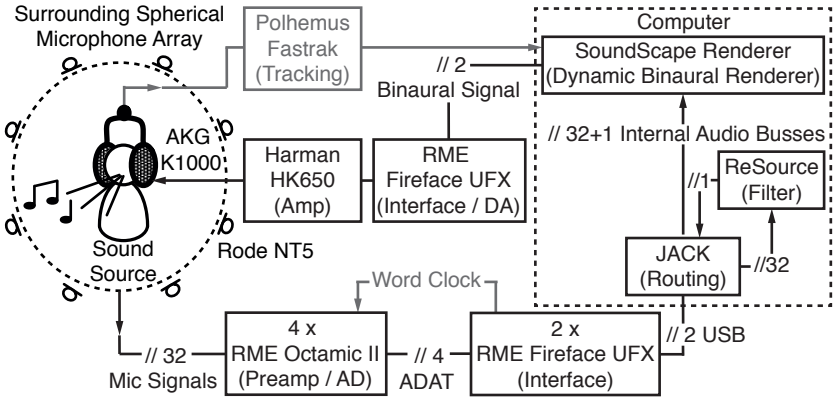

Figure 5: Setup of the reactive VAE. The pre-amplified and converted audio signals are routed with JACK to the directional-BRIR sources in the SoundScape Renderer as well as to the ReSource module. The module provides the reverberation source signal, which is passed to the diffuse-BRIR source in the renderer. The converted and amplified binaural signal is then played to the user over the AKG K1000 extra-aural headphones. A Polhemus Fastrak provides the head-tracking data.

G. LEVEL CALIBRATION

The correct level ratio between self-generated direct sound and synthesized room response is hugely important for an accurate reproduction. Thus, the entire system has to be level calibrated carefully. Of course, first of all, the levels of the microphones and BRIRs must be set to reasonable values to ensure a good overall signal-to-noise ratio. The idea of the actual calibration method 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 at the ears of a dummy head.[8,9] To conduct the calibration measurements, we placed a Genelec 1029A loudspeaker at the center of the array and a Neumann KU100 dummy head at a distance of 5 m. We then played a pink noise sequence over the loudspeaker and simultaneously recorded the sound with an array microphone located in front of the loudspeaker as well as with the dummy head. Next, we auralized this scene, using the microphone recording as the source signal and a simulated anechoic BRIR describing this scene 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 Hz, f_h = 5 kHz). This allowed for a better comparison of the RMS levels. As a results of the calibration, the real and the synthesized scene 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 RAVEN simulation remains unchanged.

4. TECHNICAL EVALUATION

A system with so many different technical components obviously has several potential sources of error. Therefore, an extensive technical evaluation is mandatory. Checking every software and hardware component in detail is one part of the evaluation, but it is also necessary to consider problems and shortcomings caused by the general design and implementation of the system. In the following, we briefly discuss some tests on basic functionality of the system as well as some systemic shortcomings.

A. BASIC FUNCTIONALITY

Considering the microphone array, there are several possible inaccuracies affecting the result of the binaural reproduction. To begin with, the directivity resolution substantially depends on the number of microphones. Given the 32 microphones, we designed the array according to the current state-of-the-art[11,20] and therefore it should be suitable for directivity measurements, especially within the scope of real-time usage where a higher spatial resolution might not even be perceivable for the user. Another important factor is the accurate positioning of the microphones. It is evident that the position of the microphones in the real construction might slightly differ from the exact position in the CAD model. However, in close collaboration with the mechanical workshop of our institute, we tried to construct the array as accurately as possible in order to comply with the required dimensions. To further ensure that the microphones are in position, we carefully measured (and adjusted) the distance between the center of the array and each microphone. Moreover, we calibrated each microphone with a Bruel & Kjaer 4230 sound level calibrator. Additionally, measuring the influence of the foam absorbers showed that interference artifacts are reduced to a negligible level. In this context, we also analyzed if it is necessary to cover the aluminum truss system and the stands with foam. However, our measurements showed no relevant influence of both items, which is why we refrained from covering them.

On the software side, we carefully examined the specially developed Matlab toolbox. To catch errors, the toolbox contains several checks and also plots every relevant processing step so that interim results can be monitored. We also conducted several tests in a shoebox-shaped test room with different image-source orders and source / receiver positions in order to check the ray-segment assignment process. As can be seen in Fig. 6 (left), we therefore displayed the image-source paths provided by RAVEN and placed a model of the array at the center of the source / receiver. This approach makes it easy to verify the results of the simulation and of the assignment process. In a next step, we generated virtual scenes for the reactive VAE based on different shoebox test conditions. For each microphone, we than conducted separate impulse response measurements of the headphone output with a loudspeaker placed in the center of the array as the source. The measured impulse responses indicate whether the reflections reach the ears at the correct time and, as a consequence, if the latency-compensation works correctly. In our case, all measurements showed only negligible time deviations ($\Delta t \approx 0.3 - 0.9\,\mathrm{ms}$), certainly caused by minimal position shifts of the loudspeaker or the microphones. Additionally, these measurements allowed to (roughly) estimate if the reflections have the right direction.

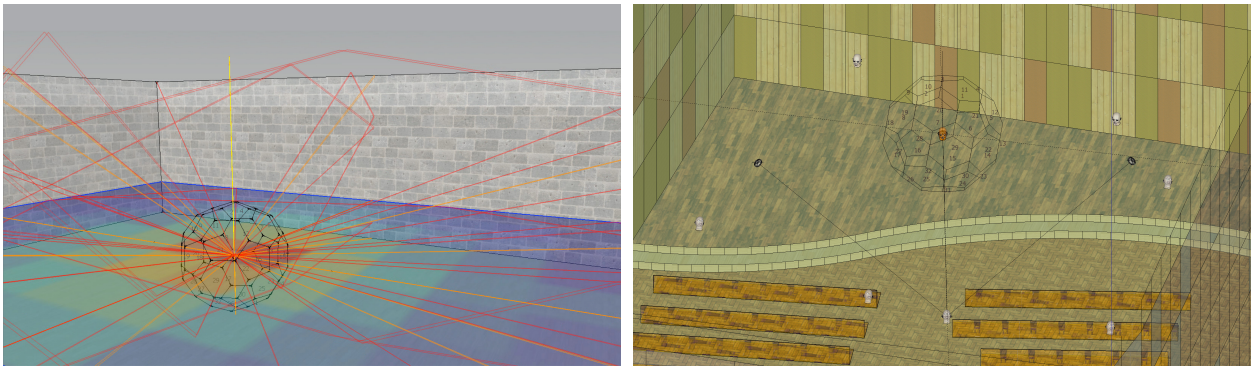

Figure 6: Sketchup models of the shoebox-shaped test room (left) and the small broadcast studio (right). The left depiction shows the image source paths (image-source order 3) and the array placed at the center of the source / receiver. This representation allows to verify the results of the simulation and of the ray-segment assignment process. The right depiction also shows the array placed at the source / receiver position on stage. This room model was used to test the entire processing chain.

Recently, we conducted similar tests with a model of the small broadcast studio (SBS) of the Westdeutscher Rundfunk (WDR). We choose this particular room for a first complete test because we have extensive measurement data of this small concert hall,[21] which we used to match the absorption coefficients of the model as well as to verify the results of the simulation and of the synthesis. Figure 6 (right) shows a section of the SketchUp model, also with the array placed at the source / receiver position on stage. Based on this model, we successfully tested the entire processing chain for different image-source orders, source / receiver positions, and reverberation modes (see section 3.C). However, these tests also showed that summing all microphone signals in time domain to obtain an excitation signal for the reverberation, as mistakenly proposed in our previous publication,[10] is not feasible. As described in section 3.D, we therefore developed the ReSource module. Unfortunately, at this point a detailed technical evaluation of the module remains to be done, but up to now the adaptive filtering works properly and provides the expected output.

Last, we evaluated the headphone and microphone compensation filters under normal use conditions. In the course of calibrating the system level, we used the KU100 dummy head to measure the direct loudspeaker signal (pink noise) as well as the binaural simulation of this scene, played back over the AKG K1000 headphones (see section 3.G). Comparing the magnitude response of both measurements indicates whether the compensation works correctly. As expected, the magnitude responses differ only slightly for frequencies above 100 Hz, which speaks for an accurate compensation. Below 100 Hz however, the anechoic chamber clearly affects the measured loudspeaker signal, but the simulation neglects the modal behavior of the room. As a result, the measurements deviate from each other in the lower frequency range.

B. SYSTEMIC SHORTCOMINGS

Besides all these factors concerning basic functionality, we also analyzed shortcomings, which inevitably come along with our general system architecture. An essential condition of the system is that the sound source is always in the center of the microphone array. Of course, this condition is not always complied, especially when the user moves. In our case, two fundamental problems occur if the acoustic center of the sound source is not in line with the center of the array. On the one hand, the sound pressure level at the microphones changes according to the 1/r law. On the other hand, the solid angles between the source and the spherical segments vary.

Regarding the level changes, a source positioned off-center results in too strong or too weak excited spherical room segments, which entails that the incident reflections per segment have the wrong level (increased or decreased according to the change in distance to the respective microphone). To approach this problem, we plan to implement another C++ real-time module which localizes the center position of the sound source via TDOA estimation (TDOA - Time Difference of Arrival), and which then adapts the gain of the microphones according to the distance to the source. This way, the reflections would always have the correct level ratio. However, in normal case, the user of the reactive VAE moves the head or rotates in the horizontal plane, but only moves slightly off center, which means that usually, the level changes due to distance shifts turn out to be relatively low. For this reason, so far we focused on the more essential components of the system (see section 3) instead of implementing the suggested adaptive gain fitting algorithm.

The change in solid angle is a problem not easy to solve. Figure 7 schematically represents this issue for one microphone (a hexagonal segment) and three different sound source positions. For a better understanding, the example shows the beam width of a cone covering the solid angle instead of the actual solid angle in steradian. As can be seen, the beam width for the hexagonal segment is about 42° if the sound source is in the center (d = 1.00 m). Now as the source comes nearer (d = 0.50 m), the beam width increases to 76°, and as the source goes away (d = 1.50 m), the beam width decreases to 29°. Of course, this example describes the issue only on a single plane for one microphone. It is clear that if a sound source moves inside the array, the solid angles of all spherical segments change with respect to the position of the source. However, when synthesizing directional BRIRs, the ray-segment assignment process is based on the assumption that the

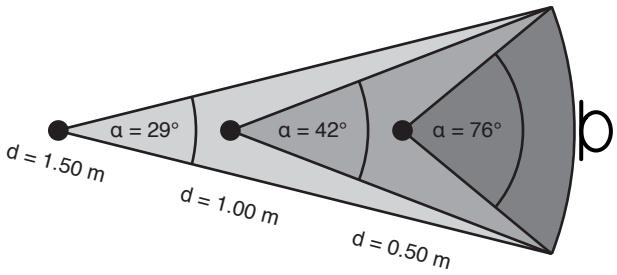

Figure 7: Schematic representation of the change in solid angle depending on the distance between the sound source and the microphone for a hexagonal segment. If the source is in the center (d = 1.00 m), the beam width for the segment is about 42°. As the source comes nearer (d = 0.50 m), the beam width increases to 76°, and as the source goes away (d = 1.50 m), the beam width decreases to 29°. The example describes the issue only on a single plane for one microphone. In real-life, the solid angles of all spherical segments change as the sound source moves.

sound source is centered (see section 3.B). Thus, the directional BRIRs are defined for fixed solid angles. As a consequence, the ray-segment assignment is only partly correct if the sound source moves and the solid angles change. More precisely, the assignment had to be adapted to the respective source position, which is not possible in our case because all directional BRIRs are pre-processed. Besides the source distance, the number of incorrectly assigned rays also highly depends on the reflection pattern of the room and on the image-source order, which is why it is hard to make a clear statement about the error. Moreover, if the assignment would be adapted, strong shifts of the sound source towards one segment would drastically decrease the directivity resolution of the system, because most of the outgoing sound rays would be assigned to the close segment with the respectively huge beam width. Either way, with the current system architecture, we cannot suggest a distinct solution approach for this issue. However, as already described in the previous paragraph, the user of the reactive VAE is mostly centered quite accurately and thus, there should be (almost) no inaccuracies in normal case. In the end, only the results of the planned psychoacoustic experiments can indicate whether these inaccuracies are perceptually relevant.

5. CONCLUSION

In this paper, we presented a system for binaural reproduction of self-generated sound in virtual acoustic environments. The so-called reactive VAE provides a new opportunity for natural interaction in the virtual room, for example by speaking or by playing an instrument. The potential applications are diverse, including use of the system as a virtual practice room for musicians, as part of an interactive virtual-reality experience, or as a research tool. In our case, the motivation for implementing the reactive VAE was primarily to be able to conduct experiments about the perception of self-generated sound as part of a recent research project.

From a technical point of view, the presented reactive VAE has two specific features: it generally works with any sound source, and it considers the dynamic directivity of the source. The key components of the system are the 32-channel surrounding spherical microphone array used to capture the user-generated sound with its particular directivity, and the synthesized BRIRs, which describe the direction-dependent reflections (directional BRIRs) as well as the binaural reverberation of the room (diffuse BRIR). These components are supplemented by the ReSource module, which provides the excitation signal for the diffuse reverberation. Also important are the compensation filters for the headphones and microphones as well as the overall level calibration to ensure the right ratio between direct sound and synthesized room response. The final setup is quite straightforward, combining the newly developed hardware and software components with available standard components.

The technical evaluation confirmed the basic functionality of the system. Within the scope of the evaluation, we performed geometric and acoustic measurements to analyze the various hardware and software components. Furthermore, using a model of a shoebox-shaped test room and of a small concert hall, we conducted extensive tests of the entire processing chain, including the ray-segment assignment process, the BRIR synthesis, and the compensation filtering. Additionally, we discussed systemic shortcomings, which mainly are the changes in level and solid angle if the acoustic center of the sound source is not in line with the center of the array. Regarding the level changes, we proposed a C++ real time module for automatic gain fitting of the microphones, which we plan to implement in future work. For the changes in solid angle, however, we could not suggest a distinct solution approach. Nevertheless, in normal case these issues should be of little consequence, because usually the user of the reactive VAE stays in the center of the array quite accurately and only moves the head or rotates in the horizontal plane.

In future work, we will refine the ReSource module in order to determine the ideal parameter settings of the algorithm, and we plan to implement the proposed C++ module for automatic gain fitting. Furthermore, we will complete our technical evaluation. As part of this, we will investigate possible issues with spatially extended sound sources. In a further perceptual evaluation, we aim to conduct listening tests to reveal the perceptual relevance of the described systemic shortcomings. Besides these aspects, we plan to examine the influence of self-generated sound on human perceptual processes by means of psychoacoustic experiments. 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.

ACKNOWLEDGMENTS

This work was funded by the German Federal Ministry of Education and Research (BMBF) under the support code 03FH014IX5-NarDasS. Many thanks to Aaron Finkenthei, Michael Söntgerath, and Marie Bauer from the mechanical workshop of our institute for their great help in designing and implementing the microphone array. We thank our student assistants Jonas Christofzik, Tim Lübeck, and Pia Krauss for their support in constructing the array. We also thank Sönke Pelzer, Lukas Aspöck, and Jonas Stienen from the Institute of Technical Acoustics (ITA) of RWTH Aachen University for their helpful correspondence.

REFERENCES

[1] S. Pelzer, L. Aspöck, D. Schröder, and M. Vorländer, "Interactive Real-Time Simulation and Auralization for Modifiable Rooms," *Building Acoustics*, 21(1), 65–74 (2014).

[2] C. Schissler and D. Manocha, "Interactive Sound Propagation and Rendering for Large Multi-Source Scenes," *ACM Transactions on Graphics (TOG)*, 36(1), 2:1–2:12 (2016).

[3] C. Pörschmann and R. S. Pellegrini, "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(11), 3–11 (2010).

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

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

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

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

[8] Z. Schärer Kalkandjiev, *The Influence of Room Acoustics on Solo Music Performances. An Empirical Investigation*. Doctoral dissertation, TU Berlin, (2015).

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

[10] J. M. Arend, P. Stade, and C. Pörschmann, "A System for Binaural Reproduction of Self-Generated Sound in VAEs," in *Proceedings of the 43rd DAGA*, 271–274 (2017).

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

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

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

[14] B. Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*. Doctoral dissertation, TU Berlin, (2016).

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

[16] P. Coleman, A. Franck, P. J. Jackson, L. Remaggi, and F. Melchior, "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc.*, 65(1/2), 66–77 (2017).

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

[18] F. Zotter, *Analysis and Synthesis of Sound-Radiation with Spherical Arrays*. Doctoral dissertation, University of Music and Performing Arts Graz, (2009).

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

[20] N. R. Shabtai, G. Behler, M. Vorländer, and S. Weinzierl, "Generation and analysis of an acoustic radiation pattern database for forty-one musical instruments," *J. Acoust. Soc. Am.*, 141(2), 1246–1256 (2017).

[21] P. Stade, B. Bernschütz, and M. Rühl, "A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios," in *Proceedings of the 27th VDT International Convention*, 1–17 (2012).