Externalization of Static Virtual Sound Sources using HRTFs Approximated by Parametric IIR Filters and Room Simulation

Patrick Nowak, Etienne Gerat, and Udo Zölzer
Helmut Schmidt University, 22043 Hamburg, Germany
Email: patrick.nowak@hsu-hh.de, e.gerat@hsu-hh.de, zoelzer@hsu-hh.de

Abstract

Externalization of virtual sound sources is still one of the major challenges in 3D spatial audio through headphones. Here, the presence of reflections and reverberation within the used directional filters, e.g. binaural room impulse responses (BRIRs), is given as the most significant factor for a successful externalization of static virtual sound sources. Since measured head-related impulse responses are usually only of short length, no room effects are included inside them. Thus, synthesized room effects have to be added afterwards in order to improve the externalization of the virtual sound sources. The image source model (ISM) and the feedback delay network (FDN) can be used to simulate early reflections and late reverberation, respectively. In this work, ISM and/or FDN are used to enhance the externalization of virtual sound sources generated using head-related transfer functions approximated by parametric infinite impulse response filters. The influence of the two room simulators on the perceived externalization of static virtual sound sources is evaluated in a listening test by comparing the achieved externalization levels to the ones achieved using measured BRIRs as directional filters.

Introduction

Nowadays, the externalization of virtual sound sources is one of the major challenges in 3D spatial audio through headphones. Here, a virtual sound source is specified as being externalized if the virtual sound source is perceived outside of the head. Room effects inside used binaural room impulse responses (BRIRs), e.g. early reflections and late reverberation, are given as the most significant factors for a successful externalization of virtual sound sources [1]. Similar to short head-related impulse responses (HRIRs), head-related transfer functions (HRTFs) approximated by parametric infinite impulse response (IIR) filter cascades contain no room effects, leading to a poor externalization perception. Thus, simulated room effects have to be added in order to enhance the perceived externalization. Here, early reflections can be simulated via image source model (ISM) [2] and late reverberation via feedback delay network (FDN) [3]. In this work, measured HRTFs are approximated by parametric IIR filter cascades consisting of two first-order shelving filters and ten second-order peak filters [4]. Additionally, room effects are simulated via ISM [2] and FDN with all-pass filters [5]. Then, a listening test is performed in order to evaluate the influence of simulated room effects on the perceived externalization of static virtual sound sources. Finally, conclusions are drawn and suggestions for further research are made.

Measured Impulse Responses

In order to evaluate the reduction in perceived externalization caused by missing room effects, BRIRs and HRIRs of the Neumann KU100 dummy-head are measured inside a room for audio listening in the Department of Signal Processing and Communication with dimensions $4.2\,\mathrm{m} \times 4.8\,\mathrm{m} \times 2.0\,\mathrm{m}$ (see Fig. 1). The center of the dummy-head and the loudspeaker are fixed to a height of 1.3 m and a distance of 2.11 m. In order to generate the excitation signal and to record by means of the dummy-head at a sampling rate of $f_s = 44.1\,\mathrm{kHz}$, an RME Fireface UCX audio interface is connected to the loudspeaker and the dummy-head. As excitation signal, an exponential sine sweep (ESS) in the frequency range between $f_{\mathrm{start}} = 55\,\mathrm{Hz}$ and $f_{\mathrm{end}} = fs/2$ is used [6]. The duration of the ESS is given as $T_{\mathrm{sweep}} = 3\,\mathrm{s}$. By rotating the dummy-head, azimuthal BRIRs are measured with a resolution of $\Delta\varphi = 15°$ and a length of 8192 samples.

Figure 1: Measurement setup.

Figure 2 shows the measured BRIR at the right ear of the dummy-head for a frontal sound source ($\varphi = 0°$). As can be seen, both direct signal and room effects are included in the measured BRIRs.

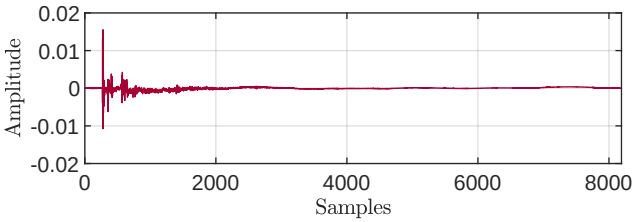

Figure 2: Measured BRIR at the right ear of the Neumann KU100 dummy-head for a frontal sound source ($\varphi = 0°$).

Afterwards, HRIRs with a length of 200 samples are extracted from the measured BRIRs by separating the im-

pulse response around the direct signal of the BRIRs. Here, the HRIRs of both ears are extracted at the same position in order to keep the interaural time difference (ITD) unchanged. Figure 3 shows the HRIR that corresponds to the BRIR from Fig. 2. Due to the small height of the room, reflections from the ceiling and floor are also included in the HRIRs in addition to the direct signal.

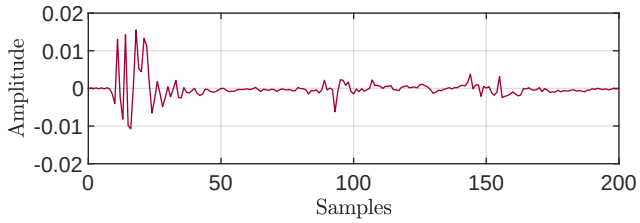

Figure 3: Shortened impulse response around the direct signal of the BRIR from Fig. 2 in order to yield an HRIR.

Then, cascades of parametric IIR filters consisting of one first-order low-frequency shelving filter, one first-order high-frequency shelving filter, and ten second-order peak filters are used to approximate the HRTF magnitude responses [4]. Figure 4 shows the approximation result for the HRIR from Fig. 3. After approximating the HRTF magnitude responses, also the ITDs have to be extracted from the HRIRs of the two ears.

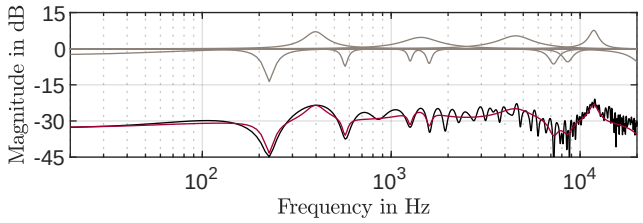

Figure 4: Approximated magnitude response (red) using a cascade of twelve parametric IIR filters (gray) for approximating the magnitude response of the HRIR from Fig. 3 (black).

Finally, the minimum-phase impulse responses can be calculated using the inverse Fourier transform (see Fig. 5). Additionally, the impulse response of the contralateral ear is delayed by the extracted ITD.

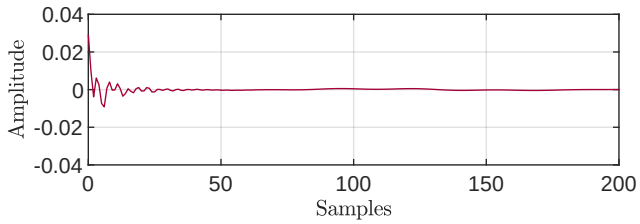

Figure 5: Corresponding minimum-phase impulse response to the approximated magnitude response from Fig. 4.

Room Simulation

Since early reflections are an important characteristic of a room, it is important to reproduce them accurately. Here, ISM is one of the most frequently used methods for simulating early reflections. The principle of ISM [2] relies on mirroring the original room at the walls to yield a number of image rooms with image sources (see Fig. 6). Afterwards, the room impulse response (RIR) can be calculated as

$$h_{\mathrm{ism}}(n) = \sum_{\mathbf{u}=0}^{1} \sum_{\mathbf{v}=-\infty}^{\infty} A(\mathbf{u},\mathbf{v}) \cdot \mathrm{sinc}(n - \tau(\mathbf{u},\mathbf{v}) \cdot f_s), \quad (1)$$

where attenuation and delay of every reflection are given as

$$A(\mathbf{u},\mathbf{v}) = \frac{\beta_{x,0}^{|v_x-u_x|}\beta_{x,1}^{|v_x|}\beta_{y,0}^{|v_y-u_y|}\beta_{y,1}^{|v_y|}\beta_{z,0}^{|v_z-u_z|}\beta_{z,1}^{|v_z|}}{4\pi d(\mathbf{u},\mathbf{v})}, \quad (2)$$

$$\tau(\mathbf{u},\mathbf{v}) = \frac{d(\mathbf{u},\mathbf{v})}{c}, \quad (3)$$

respectively. Here, c denotes the speed of sound, $d(\mathbf{u},\mathbf{v})$ the length of a reflected path, β the reflection coefficients of the individual walls, and $\mathbf{u}$ and $\mathbf{v}$ the indexing of the different image rooms.

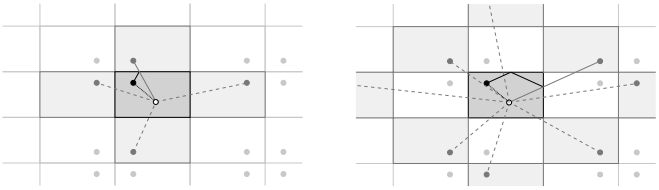

Figure 6: Principle of ISM for first- and second-order reflections in the horizontal plane.

Furthermore, when implementing the ISM, negative reflection coefficients

$$\beta = -\sqrt{1-\alpha} \quad (4)$$

should be used in order to generate reverberation tails that are similar to the characteristics of real acoustic measurements, where α denotes the absorption coefficient of the wall [7]. The absorption coefficients α are calculated according to [7] to yield a reverberation time $T_{60} = 0.2\,\mathrm{s}$ with a weighting of 0.75 for the carpeted floor and the same dimensions as the measurement room. Figure 7 shows the simulated RIR via ISM at the right ear for a frontal sound source ($\varphi = 0°$) and a length of 8192 samples, where only the direct signal is filtered with the HRTF of the corresponding direction. The RIRs of the two ears are decorrelated by spatially separating the microphone positions of the two ears by 0.2 m.

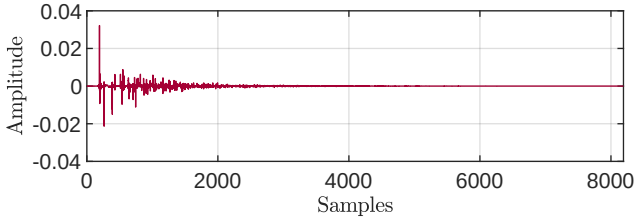

Figure 7: Simulated RIR via ISM at the right ear for a length of 8192 samples, where only the direct signal is filtered with the HRTF of the corresponding direction.

Contrarily, Fig. 8 shows the simulated RIR via ISM at the right ear for a frontal sound source ($\varphi = 0°$) only

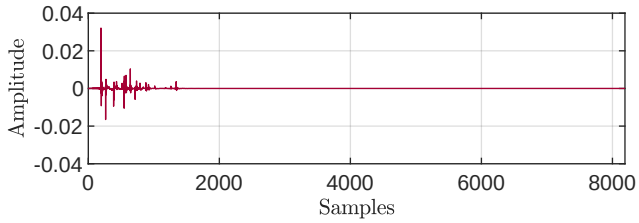

Figure 8: Simulated RIR via ISM at the right ear up to second-order reflections, where every reflection is filtered with the HRTF of the corresponding direction.

up to second-order reflections, where every reflection is filtered with the HRTF of the corresponding direction.

Since the number of image sources increases strongly with the order of reflection, the ISM is practically only used for low-order reflections. Thus, late reverberation should be simulated via another room simulator, e.g. FDN. Figure 9 illustrates the block diagram of a sparse FDN with an all-pass filter in every branch to enhance the echo density. The transfer functions of the all-pass filters are given by

$$A_k(z) = \frac{-m_k + z^{-M_k}}{1 - m_k z^{-M_k}} \qquad \text{for } 1 \leq k \leq K, \qquad (5)$$

with delay M_k and gain m_k. In this work, six parallel branches are used, where g is set to 0.2, m_k ranges from 0.63 to 0.7, and D_k and M_k are prime numbers ranging from 113 to 149. In order to decorrelate the simulated RIRs for the two ears, a_k is chosen differently for both ears between -0.15 and 0.44.

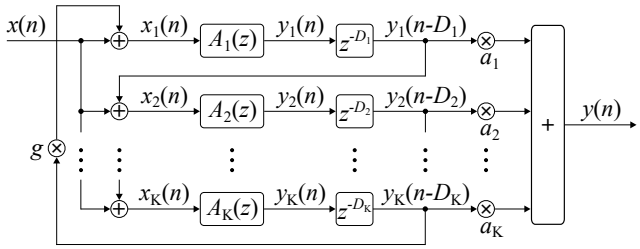

Figure 9: Block diagram of a sparse FDN with an all-pass filter in every branch.

Figure 10 shows the simulated RIR via FDN for the right ear scaled by 0.04 in order to better match the reverberation inside the measured BRIRs.

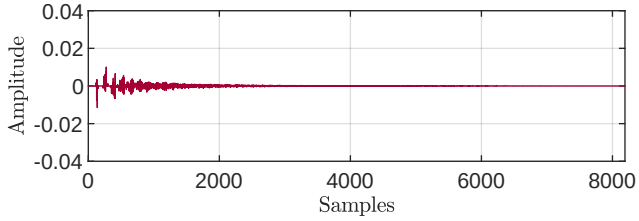

Figure 10: Simulated RIR via FDN for the right ear.

Instead of simulating a complete RIR, only late reverberation is simulated by the FDN. Thus, the direct signal and early reflections have to be appended in order to achieve a complete room simulation. In Fig. 11, the IIR filter impulse response from Fig. 5 and the RIR via FDN from Fig. 10 are added to combine direct signal and late reverberation. Here, late reverberations are delayed by 600 samples. In order to include also the early reflections, the addition of the RIR via ISM up to second-order reflections from Fig. 8 and the RIR via FDN from Fig. 10 delayed by 600 samples in comparison to the direct signal of the ISM is shown in Fig. 12. Thus, the resulting simulated RIR contains all room effects, namely direct signal, early reflections, and late reverberation.

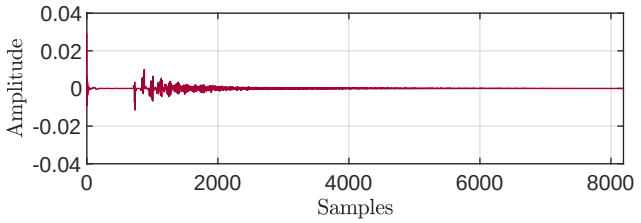

Figure 11: Combination of the IIR filter impulse response from Fig. 5 and the RIR via FDN from Fig. 10.

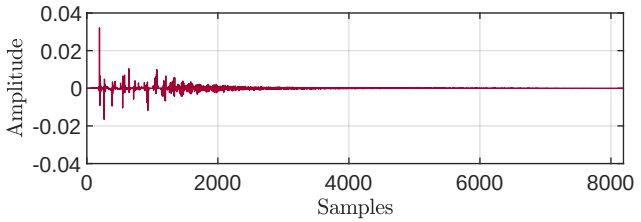

Figure 12: Combination of RIR via ISM up to second-order reflections from Fig. 8 and RIR via FDN from Fig. 10.

Listening Test

In order to evaluate the externalization capabilities of the different filter types, a listening test is performed in which the subjects rate the perceived externalization of different stimuli for given azimuthal directions in between 0 and 1, where a value of 0.25 defines the surface of the head such that lower values signify internalized virtual sound sources. As stimulus, either a single snap or a single word ("*matter*") are used. The stimuli can be played as many times as desired before submitting the perceived externalization. Additionally, a text box allows to enter a comment. Seven different filter types are included in the listening test, namely BRIR (Fig. 2), HRIR (Fig. 3), IIR (Fig. 5), ISMfull (Fig. 7), ISM2nd (Fig. 8), FDN (Fig. 11), and ISMandFDN (Fig. 12). For every filter type six different azimuthal directions are evaluated ($\varphi = -90°, -45°, 0°, 90°, 135°, 180°$), resulting in a total of 84 stimuli per test run and a duration of approximately 15 minutes. Every subject is asked to perform two test runs with a short break in between. Overall, twelve subjects participated in the listening test using the Beyerdynamic DT770 Pro 250 Ohm over-ear headphone. The participants were eleven men and one woman, aged between 27 and 64 years with an average age of 33.9 years.

Table 1 summarizes the average perceived externalization per filter type and subject. As can be seen, the average perceived externalization varies strongly with subject. However, ISMandFDN reaches the highest average

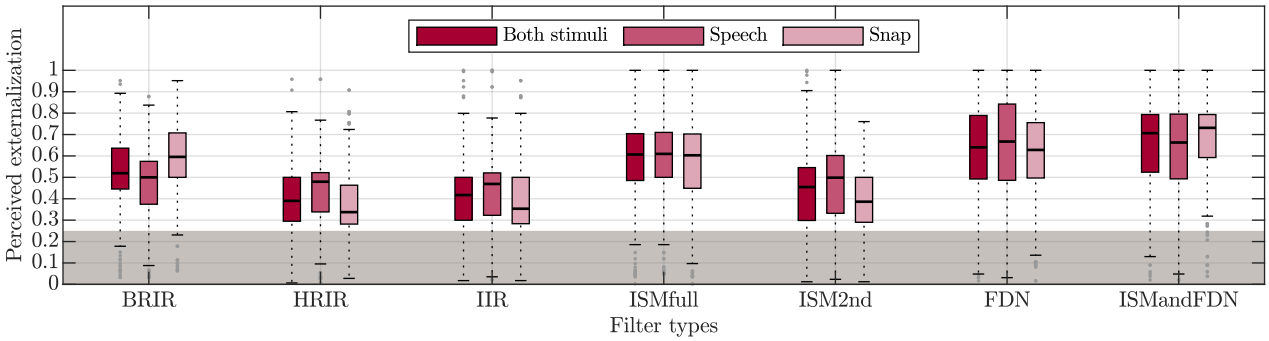

Figure 13: Box-and-whisker plots for the different filter types containing lower whisker, first quartile, median, third quartile, upper whisker, and outliers. Additionally, the results are separated for the different stimuli.

Table 1: Average perceived externalization separated for the different filter types and subjects. The values are represented in different colors in order to highlight the **maximum** value per subject and values reaching at least 90 % or 70 % of the maximum value. Values below this threshold are represented in light red.

Subject	1	2	3	4	5	6	7	8	9	10	11	12	All
BRIR	0.500	0.567	0.603	0.538	0.456	0.267	0.549	**0.471**	0.561	0.563	0.560	0.641	0.523
HRIR	0.336	0.397	0.324	0.459	0.439	0.274	0.478	0.334	0.410	0.437	0.453	0.516	0.405
IIR	0.393	0.419	0.312	0.484	0.499	0.287	0.415	0.367	0.386	0.411	0.446	0.578	0.416
ISMfull	0.595	0.648	0.652	0.535	0.504	0.273	0.530	0.460	**0.593**	**0.630**	0.666	0.778	0.572
ISM2nd	0.424	0.515	0.378	0.403	0.405	0.177	0.406	0.249	0.492	0.494	0.515	0.689	0.429
FDN	0.692	0.836	0.691	0.661	**0.703**	**0.313**	0.608	0.339	0.549	0.534	0.678	**0.934**	0.628
ISMandFDN	**0.772**	**0.858**	**0.742**	**0.732**	0.684	0.262	**0.633**	0.464	0.579	0.594	**0.721**	0.905	**0.662**

perceived externalization for six out of twelve subjects. Additionally, FDN, ISMfull, and BRIR, achieve the maximum average perceived externalization for three, two, and one subject, respectively. Similar results are found in the box-and-whisker plots in Fig. 13. Due to missing room effects, HRIR and IIR have the lowest perceived externalization whereas FDN and ISMandFDN have the highest perceived externalization without an overlap of the interquartile range with HRIR and IIR. Here, FDN and ISMandFDN even clearly outperform BRIR, which according to comments can be explained by stronger room effects than in the measured BRIRs. Contrarily, adding only early reflections (ISM2nd) has a very small effect on perceived externalization. An evaluation across the different azimuthal directions has shown that lateral sound sources externalize better than frontal or rear sound sources, which is explained by the position of the headphone's loudspeakers.

Conclusion

Binaural synthesis using measured BRIRs is able to generate externalized static virtual sound sources. However, when using shortened impulse responses, like HRIRs or HRTFs approximated by parametric IIR filter cascades, the perceived externalization is strongly reduced due to missing room effects. Adding synthesized room effects via ISM and FDN increases the perceived externalization even above the level achieved by measured BRIRs. Here, reverberation is more important than early reflections. In future work, fine-tuning of room simulation parameters has to be done in order to better match measured impulse responses and reduce audible artifacts. Additionally, the influence of synthesized room effects on the perceived direction has to be evaluated.

References

[1] D. R. Begault, A. S. Lee, E. M. Wenzel, M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source", *in Audio Engineering Society Convention 108*, 2020.

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics", *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[3] J.-M. Jot and A. Chaigne, "Digital delay networks for designing artificial reverberators", *in Audio Engineering Society Convention 94*, 1991.

[4] P. Nowak, "Spatial Audio Through Headphones Based on HRTFs Approximated by Parametric IIR Filters", *PhD thesis, Helmut Schmidt Universität, Hamburg*, 2022.

[5] U. Zölzer et al., *Digital Audio Signal Processing*, John Wiley & Sons Ltd, Chichester, 3rd edition, 2022.

[6] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique", *in Audio Engineering Society Convention 108*, 2000.

[7] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model", *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.