Investigating Higher Order Spatial Impulse Response Rendering for the evaluation of hearing devices

Laurent S. R. Simon[1,2], Hannes Wüthrich[2], Norbert Dillier[1]
[1] *University of Zürich (UZH), Zürich, Switzerland*
[2] *Sonova AG, Stäfa, Switzerland, Email: laurent.simon@sonova.com*

Introduction

In recent years, more and more perceptual studies involving hearing devices have been conducted using virtual sound environments (VSE). Using VSEs leads to different - and presumably more realistic - speech intelligibility measurements [1, 2], and makes it also possible to study localization [3, 4, 5], auditory distance perception, behaviour in noisy conditions [6], and other attributes in more ecologically-valid environments. Among the existing 3D sound reproduction techniques, higher order Ambisonics (HOA) is one that can be used both to record and reproduce plausible sounds, either on loudspeakers or on headphones [7, 8]. At high frequencies, however, the sound field produced by HOA suffers from spatial aliasing, which alters interaural level differences (ILD). Outside of the sweet spot, i.e. the center of the sound reproduction loudspeaker array, the spatial aliasing increases and can also alter the interaural time differences (ITD). In hearing research, several studies showed that the artifacts produced by HOA may lower the performance of beamformers or of other hearing device algorithms [9, 10, 11, 12].

In this paper, we investigate the use of spatial decomposition of impulse responses and rendering via a mix of HOA and either vector-base amplitude panning (VBAP) or nearest speaker panning (NSP), as a way to to ensure realistic beamformer outputs and interaural level differences outside of the sweet spot. The main spatial decomposition algorithm tested in this study was the higher order spatial impulse response rendering (HO-SIRR) [13]. It aims to calculate a number of HOA analysis beams. In each of those beams, it performs a time-frequency analysis, and in each time-frequency bin, the ratio between the energy vector and the intensity vector is used to deduce a ratio of directionality. The directional parts of the time-frequency bins are panned on the reproduction loudspeaker setup using VBAP, whereas an HOA impulse response is created with the diffuse components. The diffuse HOA RIR is then decoded for the loudspeaker setup. The resulting diffuse loudspeaker signals are subsequently decorrelated. An alternative HO-SIRR that replaces the VBAP panning of directional components with NSP was also tested, AHO-SIRR.

Impulse responses were recorded with a KEMAR manikin wearing dummy hearing aids, as well as with an Eigenmike microphone. Recordings were made in three different rooms and HOA, HO-SIRR, AHO-SIRR, and the KEMAR reference were evaluated in two different reproduction rooms.

Experimental conditions

Room impulse responses (RIR) were recorded in three different rooms: a dry space (RT60 = 222ms, DRR = 7.7dB), an office space (RT60 = 400ms, DRR = 4.37 dB), and a hall (RT60 = 866ms, DRR = 8.3dB) using 5s-long exponential sweeps ranging from 100Hz to 20kHz [14]. In each of these rooms, 12 8020 Genelec loudspeakers were placed every 30° at a height of 1.20m. The RIR were measured either using both an Eigenmike EM32 microphone at the center of the array or a KEMAR wearing hearing aid shells that contained hearing aid microphones. Recordings with the KEMAR were conducted at the center of the loudspeaker array (0cm condition), 20cm off-centered, and 50cm off-centered.

The KEMAR recordings were used as a reference.

The reproduction techniques being evaluated were HOA, HO-SIRR, and the alternative HO-SIRR (AHO-SIRR). The 4th order Ambisonics RIRs were decoded using the Higher Order Ambisonics Library [15]. A single band energy-preserving decoder with $\max_{RE}$ optimization was used. The HO-SIRR was calculated using multiresolution short-time Fourier Transform (512 samples overlapping windows until 500 Hz, 128 samples overlapping windows between 500Hz and 2kHz, and 64 samples overlapping windows above 2kHz). The diffuse rendering option was activated, the broadband analysis was deactivated, and the diffuseness parameter temporal averaging coefficient was equal to 0.5. The AHO-SIRR was calculated using the same parameters, but the vector-base amplitude panning (VBAP) coefficients were replaced with those of NSP.

The three tested techniques were decoded for two different rooms. Room A was a small room with a reverberation time of 117ms and an extended hemispherical array of 32 loudspeakers, similar to the one described in [11]. Room B was a large room (193 m^3), the dry space described above. It contains 105 loudspeakers placed all around the room [16]. In this room, participants are encouraged to move around and therefore are rarely standing at the sweet spot. It was therefore important to study the sound field reproduction outside of the sweet spot. For the HOA reproduction, a subset of 30 loudspeakers was used, chosen to be the closest loudspeakers to those of a t-design [17] loudspeaker array of 36 loudspeakers. For the HO-SIRR, however, although the diffuse part of the RIR would remain in HOA and would therefore be decoded for the same subset of 30 loudspeakers as HOA, the directional part could in theory benefit from having

more loudspeakers. 16 out of the 105 loudspeakers of room B were dismissed, as their direction and directivity could not ensure the same frequency response as that of the other loudspeakers. In room B, the directional part of the HO-SIRR RIR was therefore decoded over the 89 remaining loudspeakers. A version of HO-SIRR using the same set of 30 loudspeakers for both the diffuse and directional part was also tested. For the analysis of the results, in room A, HOA and HO-SIRR were compared to AHO-SIRR, whereas in room B, HOA and the HO-SIRR using 89 loudspeakers for the directional sounds were compared to the HO-SIRR using 30 loudspeakers for both the directional and diffuse part of HO-SIRR.

The various tested techniques were compared in terms of front-back ratio (FBR) of a static monaural beamformer. This is a measure of how directive the beamformer is at every third-octave band between 160Hz and 8kHz.

The ILDs were calculated in 20 ERB frequency bands, between 50 Hz and 8.5 kHz, for each RIR direction. An average difference of ILD between each tested technique and the reference was computed by averaging the absolute error of ILD for all frequency bins between 2 and 6kHz, for all positions, ILD being mostly used in this frequency range. The difference of ILD between each tested technique and the reference was then computed.

Finally, the ITDs were estimated using the cross-correlation of the envelope of the broadband impulse responses [18].

Results

Fig. 1 to 3 display some examples of FBR measurements. All the FBR measurements follow the same trend, i.e. the FBR of HO-SIRR and AHO-SIRR are closer to the reference FBR than the FBR of HOA.

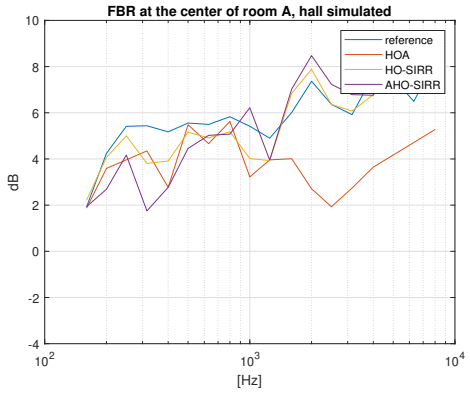

Figure 1: FBR of the reference, HOA, HO-SIRR and AHO-SIRR when the hall is reproduced in room A at the center of the loudspeaker array.

The average ILD errors are shown in Table 1 and 2.

Similarly to the FBR, several examples of ITD comparisons are shown in Fig. 4 to 6. However, these show that the ITD estimated with the cross-correlation of the envelope of the impulse response is noisy. As an alternative, both the cross-correlation estimator and the the threshold estimator were tested for various threshold values on

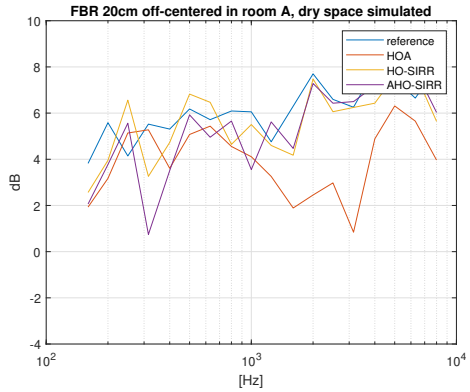

Figure 2: FBR of the reference, HOA, HO-SIRR and AHO-SIRR when the dry space is reproduced in room A 20cm off-centered.

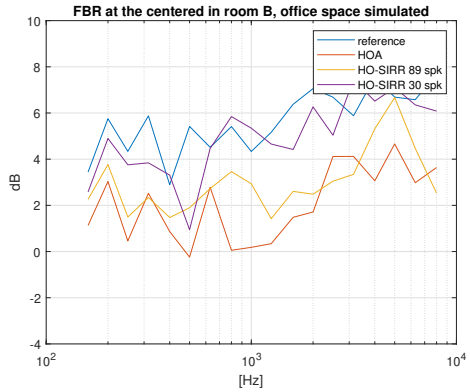

Figure 3: FBR of the reference, HOA, HO-SIRR with 89 loudspeakers for the directional sound, and HO-SIRR with 30 loudspeakers when the office space is reproduced in room B at the center of the loudspeaker array.

the low-pass filtered impulse responses, but the results were similarly bad for the tested techniques outside of the sweet spot. All of these estimators gave good results for synthesized anechoic HOA impulse responses, as can be seen in Fig. 7. It is therefore the authors' assumption that both the reverberation and the errors of reconstruction make these estimations more fragile and less reliable. A further study would be required to better understand these observations.

Discussion

In room A, HO-SIRR systematically outperformed HOA with regards to the FBR, especially at high frequencies, where the spatial aliasing of HOA was less of a problem for the VBAP directional parts of the impulse responses and for the decorrelated diffuse sound. It also outperformed HOA in terms of ILD errors. In room B, using 89 loudspeakers for the reproduction of the directional part of HO-SIRR led to similar results to using the same 30 loudspeakers for both components of HO-SIRR. Both showed a significant performance improvement compared to HOA. When simulating the office space in room B, the performance of the FBR was lower when using 89 loudspeakers than when using 30 loudspeakers. An informal listening of the HOA and HO-SIRR in that condition

Table 1: ILD errors between the tested techniques and the reference, averaged between 2 and 8kHz (in dB), in room A

Room and distance	HOA	HO-SIRR	AHO-SIRR
hall 0cm	3.9	2.3	2.0
hall 20cm	5.0	2.9	2.9
hall 50cm	3.9	2.2	2.1
office 0cm	3.6	2.3	1.9
office 20cm	4.0	2.2	2.2
office 50cm	3.8	2.2	2.1
dry 0cm	3.6	1.8	1.7
dry 20cm	4.0	1.9	1.8
dry 50cm	3.6	1.9	1.9

Table 2: ILD errors between the tested techniques and the reference, averaged between 2 and 8kHz (in dB), in room B

Room and distance	HOA	HO-SIRR 89	HO-SIRR 30
hall 0cm	5.2	3.9	3.6
hall 20cm	4.7	3.9	4.1
hall 50cm	4.2	3.5	3.2
office 0cm	4.4	3.4	3.2
office 20cm	4.1	3.7	3.3
office 50cm	3.8	3.4	3.0
dry 0cm	5.1	3.1	3.1
dry 20cm	4.4	3.4	3.3
dry 50cm	3.8	3.2	2.9

showed that the direction of the sound in the original HOA recording was hardly perceptible, possibly because of the low DRR, in which case the lower number of loudspeakers in the 30 loudspeakers condition could have reduced the errors of direction of arrival estimations of HO-SIRR.

The estimated ITDs displayed more errors than expected. The estimation of ITDs via loudspeaker sound reproduction is in any case biased by the ITD of each of the loudspeaker, independently of the theoretical ITD of the source. The threshold method - with thresholds varying between -18 and -3dB - and the cross-correlation of the envelope of the low-pass filtered RIRs - with a cut-off frequency of 1.5kHz - were also tested but led to worse estimations of the ITDs. Given that in [11], the estimated ITDs were closer to those of the reference, and since the main experimental differences were the different reproduction room and the fact that in the article cited above, the simulated impulse responses were synthesized and anechoic, it is thought that the main reason for the observed differences was the reverberation of the rooms. The interaction between the reverberation and the artifacts of HOA and HO-SIRR is also a possible cause for these poor ITD estimations.

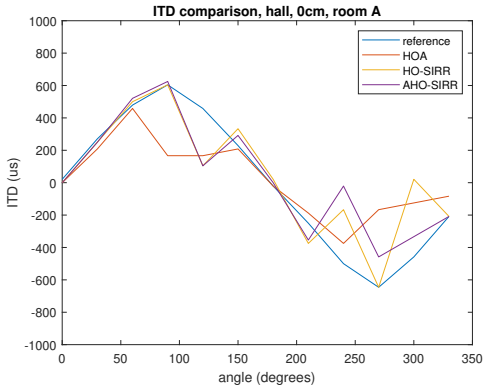

Figure 4: ITD of the reference, HOA, HO-SIRR and AHO-SIRR when the hall is reproduced in room A at the center of the loudspeaker array.

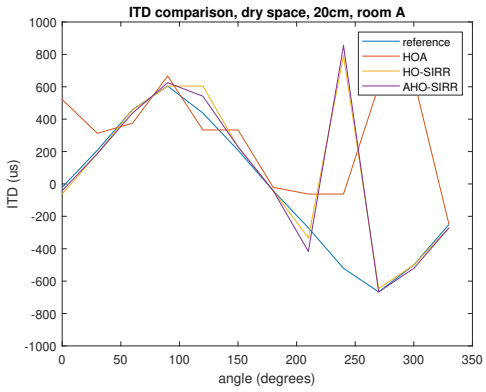

Figure 5: ITD of the reference, HOA, HO-SIRR and AHO-SIRR when the dry space is reproduced in room A 20cm off-centered.

Conclusion

HO-SIRR showed a systematic and significant improvement over HOA with RIR recorded in three different rooms and reproduced on a 32-loudspeakers setup, in terms of FBR and ILD.

In a larger room with a longer reverberation time and more loudspeakers, using the same loudspeakers for HO-SIRR and HOA, HO-SIRR performed better than HOA. Using a larger number of loudspeakers for the directional part of HO-SIRR did not lead to a significant improvement of the ITD and FBR in those rooms.

The alternative HO-SIRR that replaced VBAP with NSP showed similar results to those of HO-SIRR. An improvement over HO-SIRR was observed for the ILD estimations.

The ITD estimations of HOA and HO-SIRR were noisy, presumably because of the combination of the room reflections and of the reproduction techniques artifacts. The estimated ITD was however better with HO-SIRR than with HOA.

References

[1] Mansour, N., Marschall, M., Westermann, A., May, T., and Dau, T: Speech Intelligibility in a Realis-

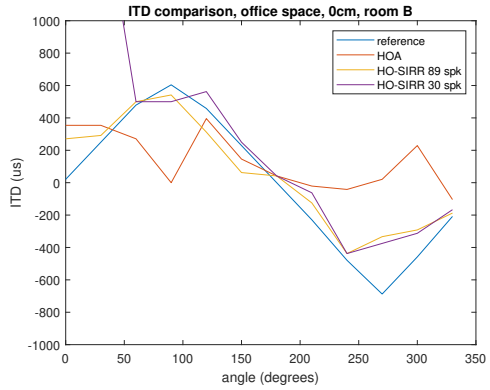

Figure 6: ITD of the reference, HOA, HO-SIRR with 89 loudspeakers for the directional sound, and HO-SIRR with 30 loudspeakers when the office space is reproduced in room B at the center of the loudspeaker array.

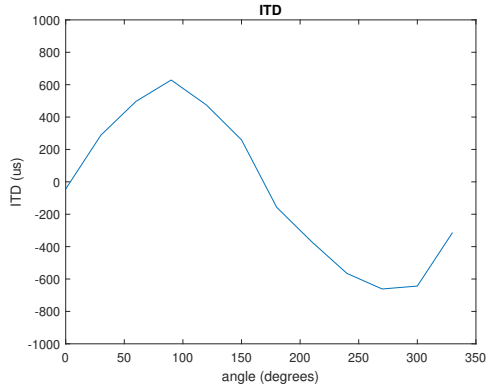

Figure 7: Example of ITD estimation for synthesized HOA impulse responses simulated on a KEMAR manikin in an anechoic room, using the threshold method on the low-pass filtered impulse response and a threshold of -10dB

tic Virtual Sound Environment, 23rd International Congress on Acoustics, 2019

[2] Badajoz-Davila, J. and Buchholz, J: M.: Effect of Test Realism on Speech-in-noise Outcomes in Bilateral Cochlear Implant Users, Ear and Hearing, 2021

[3] Simon, L. S. R., Kegel, A., Wüthrich, H., Dillier, N.: 3D localization of speech by mildly and moderately hearing-impaired persons in ecological environments, 23rd International Congress on Acoustics, 2019

[4] Müller, M. F., Kegel, A., Schimmel, S., Dillier, N,, Hofbauer, M.: Localization of virtual sound sources with bilateral hearing aids in realistic acoustical scenes, The Journal of the Acoustical Society of America, vol. 131, issue 6, 2012

[5] Baumgartner, R., Madjak, P., Laback, B.: Modeling sound-source localization in sagittal planes for human listeners, The Journal of the Acoustical Society of America, vol. 136, issue 2, 2014

[6] Weisser, A., Miles, K., Richardson, M. J., and Buchholz, J. M.: Conversational distance adaptation in noise and its effect on signal-to-noise ratio in realistic listening environments, The Journal of the Acoustical Society of America, vol. 149, issue 4, 2021

[7] Daniel, J., Moreau, J., Nicol, R.: Further Investigations of High-Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging, 114th Audio Engineering Society Convention, 2003

[8] Zotter, F., Pomberger, H., Noisternig, M.: Ambisonic Decoding with and without Mode-Matching: A Case Study Using the Hemisphere, 2nd Int. Symposium on Ambisonics and Spherical Acoustics, 2010

[9] Oreinos, C. and Buchholz, J. M.: Evaluation of Loudspeaker-Based Virtual Sound Environments for Testing Directional Hearing Aids, Journal of the American Academy of Audiology, vol. 27, issue 7, 2016

[10] Grimm, G., Ewert, S., Hohmann, V.: Evaluation of Spatial Audio Reproduction Schemes for Application in Hearing Aid Research, Acta Acustica united with Acustica, vol. 101, issue 4, 2015

[11] Simon, L. S. R., Wüthrich, H., Dillier, N.: Comparison of 3D Audio Reproduction Methods Using Hearing Devices, Journal of the Audio Engineering Society, vol. 68, issue 12, 2020

[12] Giurda, R., Simon, L. S. R., Wüthrich, H., Dillier, N.: Evaluation of an ILD-based hearing device algorithm using Virtual Sound Environments, 23rd International Congress on Acoustics, 2019

[13] McCormack, L., Pullki, V., Politis, A., Scheuregger, O., Marschall, M.: Higher-Order Spatial Impulse Response Rendering: Investigating the Perceived Effects of Spherical Order, Dedicated Diffuse Rendering, and Frequency Resolution, Journal of the Audio Engineering Society, vol. 68, issue 5, 2020

[14] Farina, A., Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique, 108th Audio Engineering Society convention, 2000

[15] Politis, A., Microphone array processing for parametric spatial audio techniques, PhD thesis, Aalto University, 2016

[16] Klockgether, S.: The freedom to move around - Introduction of the Sonova real life lab, The 179th Meeting of the Acoustical Society of America, 2020

[17] Hardin, R. H. and Sloane, N. J. A.: McLaren's improved snub cube and other new spherical designs in three dimensions, Discrete & Computational Geometry, vol. 15, 1996

[18] Katz, B. F. G. and Noisternig, M.: A comparative study of interaural time delay estimation methods, Journal of the Acoustical Society of America, vol. 135, issue 6, 2014