Perceptual Evaluation of Spatial Resolution in Early Reflections

Matthias Frank, Manuel Brandner, Franz Zotter

Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria

Emails: frank@iem.at, brandner@iem.at, zotter@iem.at

Introduction

Real-time auralization of rooms generally puts a sound source into a virtual room by convolving the dry source signal with a spatial room impulse response to create the appropriate binaural headphone signals. Nowadays, auralization at least allows the listener to rotate the head (3 degrees of freedom, 3DoF) or even to walking around in the virtual room (6 degrees of freedom, 6DoF). Spatial room impulse responses are either created data-based, i.e. by measurement of existing rooms or by calculating detailed room simulations, or model-based, i.e. by simplified room models. A disadvantage of the first type is the extensive measurement or simulation effort and the amount of data, especially for 6DoF applications. In addition, the artifact-free interpolation between different listening positions is not an easy task [1]. This is much easier using simplified models that can efficiently recalculate in real time, such as image-source models of shoe-box rooms for early reflections and feedback-delay networks for late, diffuse reverberation. However, the modeling of real-world rooms, i.e. finding suitable parameters, is challenging using the before-mentioned simplifications.

Rotation for the incorporation of head orientations is mostly done by using a spherical harmonic representation [2] of the sound field around the listener's head. This makes interpolation much easier in comparison to other methods [3, 4]. This is why Ambisonics is typically used in 3DoF applications, such as 360° videos. Another advantage of Ambisonics is the scalability of the spatial resolution in terms of the maximum order to adjust the playback to available computational resources. This is especially important for interactive auralization with low latencies on mobile devices.

There has been quite some research about the required spatial resolution for authentic, i.e. not distinguishable from the real reference in a direct comparison, or plausible auralization, i.e. consistent with an inner reference [5]. Engel [6] compared order-reduced versions to a 4^{th}-order microphone measurement and found orders of 2 or 3 sufficient. Zaunschirm found orders below 5 or 7 distinguishable from a direct measurement with a dummy head [7]. In a comparison of a real loudspeaker in a real room, a 7^{th}-order simplified room simulation was perceived differently in experiments by Enge [8], however orders 3 and 7 were only different in a 3DoF scenario, but not in 6DoF, indicating a reduced sensitivity for spatial resolution when the listener can walk around in the virtual room.

Taking a closer look at the different parts of an impulse response provides the possibility of rendering the parts in different spatial resolutions to save computational effort:

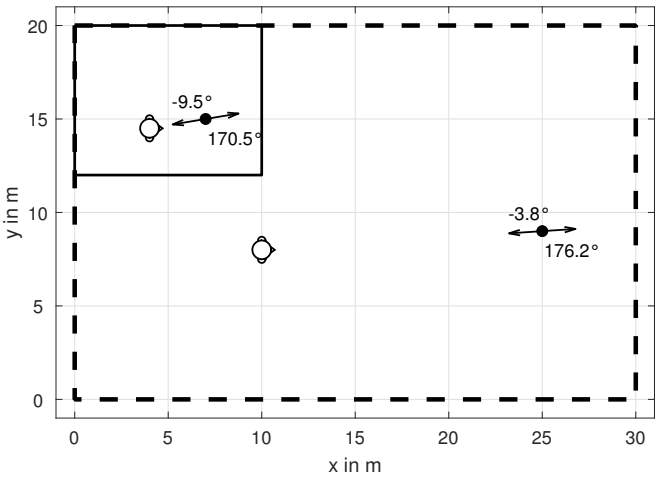


Figure 1: Listener/source position and source orientation in the horizontal cross section of the simulated rooms.

direct sound, early reflections, and late reverberation after the mixing time [9]. Direct sound is most sensitive and requires the highest resolutions. Orders up to 11, 18 [10], or even 30 [11] can be found in literature. However, the required orders could be reduced to 7 or even 3 for speech when using state-of-the-art approaches, such as magnitude least squares (magLS) by Zaunschirm and Schörkhuber [12, 13]. The study by Lübeck [10] reported orders between 3 and 12 for early reflections, and 3 to 6 for late reverberation, both using measured impulse responses without magLS. Even lower, a recent study showed that magLS is capable of perceptually recreating simulated diffuse reverberation already with 1^{st}-order resolution due to the covariance filters [14].

This contribution tries to complete the row of studies using a model-based auralization with the state-of-the-art magLS approach by evaluating the minimum required spatial resolution of early reflections using an image-source model of a simple shoe-box room. While the resolution of the direct sound is kept at 7^{th} order for all conditions, the early reflections were played back in orders 0 to 6 and are compared to a 7^{th}-order reference. The comparisons are done for speech and noise, different number of reflections, different source orientations, and different sizes of the virtual room.

The paper first describes the details of the experimental setup and the evaluated conditions. Then, the results are presented and discussed with regard to possible explanations. Finally, the paper is summarized and new questions for subsequent research are posed.

Setup and Conditions

The room simulation employed an image-source model of a shoe-box room, as implemented in the IEM `RoomEncoder` VST plug-in[1]. Headphone playback employed head-tracked [15] binaural Ambisonics with the magLS approach [12, 13] in the `BinauralDecoder` plug-in. The differently large rooms were simulated to evaluate the influence of room size and reverberation time. The small room had a size of 10 m × 8 m × 3 m and a reverberation time of 0.6 s between 200 Hz and 2 kHz. The virtual listener was positioned at (-1, -1.5, -0.2) m and the source at (2, -1, 0.2) m relative to the center of the room, resulting in a source/receiver distance of 3.1 m, see Figure 1. The large room had a size of 30 m × 20 m × 10 m and a reverberation time of 1.9 s between 200 Hz and 2 kHz, similar as in [16, 17]. In this room, the listener was positioned at (-5, -2, -3) m and the source at (10, -1, -1) m, resulting in a source/receiver distance of 15.2 m.

The number of simulated image sources was varied between 6 (1st-order image sources) and 236 (7th order). Moreover, the source was oriented in two different ways to face completely towards the listeners or away, see Figure 1 for exact angles. In order to provoke a strong acoustic effect of the changing source orientation, the source was modeled with a first-order cardioid directivity, so that there was no direct sound at all when the source was facing away from the listener.

The experiment employed two different sounds: (a) continuous pink noise for maximum sensitivity to coloration and (b) male English speech [18] that facilitates better spatial perception and familiarity. The direct sound was always simulated with 7th-order spatial resolution, while the early reflections were simulated with varying resolution between 0 and 6. In the reference case, the entire simulation was done in 7th order.

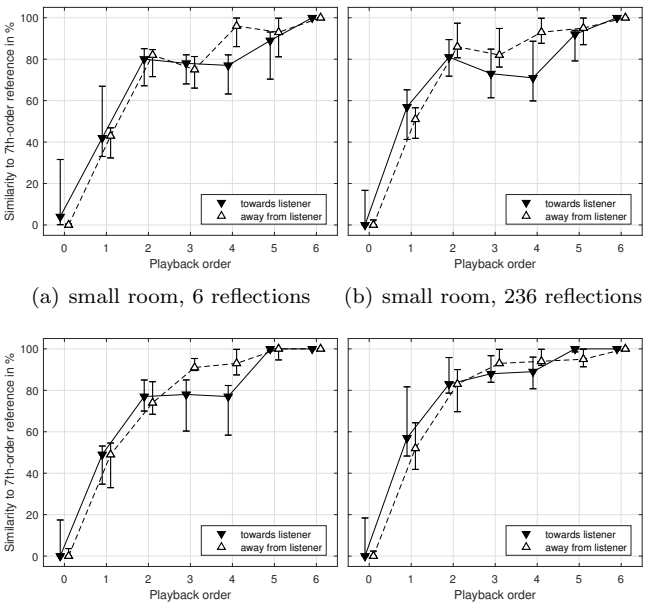
Overall, there were 16 = 2 (sounds) × 2 (rooms) × 2 (number of reflections) × 2 (source orientations) trials with multi-stimulus comparisons. The listeners' task was to compare the similarity of the 7 (0th to 6th-order resolution for the early reflections) stimuli to the corresponding 7th-order reference on a continuous scale from *very different* to *identical*. The overall playback level of each trial was manually balanced by the authors, as the different acoustical conditions yielded different loudness.

Results

On average, each of the 13 experienced listeners (1 female, 12 male) needed 27 minutes to perform the entire experiment. No data was excluded from the analysis.

Figure 2 shows the resulting median values and corresponding 95% confidence intervals for all speech conditions in dependence of room size and number of reflections. Within each sub-figure, the different markers indicate the results for the source facing towards the listener and away, respectively. It can be clearly seen that the similarity to the reference increases with the order, however, monotony is not strictly preserved for the small room. For orders around 4, the similarity is greater when the source is facing away from the listener.

[1] freely available at `plugins.iem.at`



(a) small room, 6 reflections (b) small room, 236 reflections

(c) large room, 6 reflections (d) large room, 236 reflections

Figure 2: Medians and 95% confidence intervals of perceived similarity to 7th-order reference for noise.

Table 1: Minimum required orders for noise to be indistinguishable from 7th-order reference at 5% level with Bonferroni-Holm correction.

	small room		large room	
	6	236	6	236
towards listener	6	6	5	5
away from listener	6	6	5	6

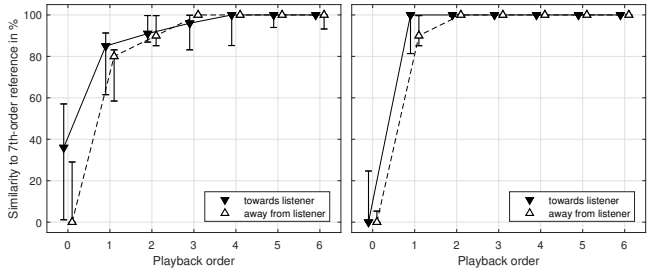
In order to statistically determine the perceptually required order, a Bonferroni-Holm-corrected Wilcoxon signed-rank test was carried out between the results for each order and the reference. The minimum required order was then defined as the lowest order that yielded a p-value ≥ 0.05, i.e. the first order that is not significantly different from the reference. The resulting minimum required orders for noise are summarized in Table 1. While the small room required orders of 6, the larger room was less sensitive, except in the case of the source facing away using 236 reflections. There was no general influence of the number of reflections on the minimum required order.

As the values from Table 1 did not reflect the visible trend of greater similarity for sources facing away from the listener, an alternative definition of the minimum required order was designed. This time, it was defined as the minimum order that achieved a median similarity to the reference of at least 90%. The values in Table 2 indicate that the required orders did not really change for the source facing towards the listener, however, when facing away, the values are reduced by 2 orders, except for the large room with 236 reflections. While with the alternative definition of the minimum required order, there was no general influence of the number of reflections, the larger room required less resolution.

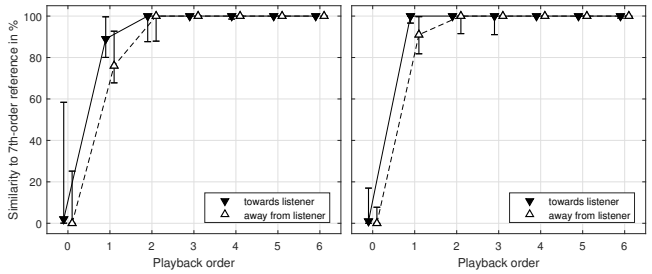
Table 2: Alternative minimum required orders for noise to achieve a median similarity of at least 90%.

	small room		large room	
	6	236	6	236
towards listener	6	5	5	5
away from listener	4	4	3	3

The increased sensitivity when the source is facing the listener could be explained by the shorter distance between the two first active sound paths and hence stronger combfilters: Between direct sound and floor reflection, the delay was only 1 ms and 3 ms for the small and the large room, respectively. In contrast, when the source was facing away, the delay between the floor reflection and the front wall was 6 ms and 7 ms. Moreover, the interaction between the different-order direct sound and first reflection could cause additional coloration. In the used magLS decoder, the crossover frequency, above which the phase of the head-related impulse is neglected, is order-dependent. Mixing two sound paths with different crossover frequencies and small delay might lead to audible coloration. It would be interesting to compare the results to a processing that uses the same reduced order for both direct sound and reflections.



(a) small room, 6 reflections (b) small room, 236 reflections



(c) large room, 6 reflections (d) large room, 236 reflections

Figure 3: Medians and 95% confidence intervals of perceived similarity to 7^{th}-order reference for speech.

For speech, the results were generally closer to the reference, cf. Figure 3. Here, the similarity monotonically increased with the order. The minimum required orders as shown in Table 3 reveal an order of 3 for the small room and an order of 2 for the large room, both with only 6 reflections. When rendering 236 reflections, sensitivity decreased towards first order, except for the small room when the source is facing away from the listener.

There was no general influence of the source orientation, however, sensitivity decreased with both the room size and the number of reflections. As in practice, more than 6 reflections are typically rendered, orders of 1 or 2 seem to be sufficient for plausible playback of early reflections generated by an image-source model.

Table 3: Minimum required orders for speech to be indistinguishable from 7^{th}-order reference at 5% level with Bonferroni-Holm correction.

	small room		large room	
	6	236	6	236
towards listener	3	1	2	1
away from listener	3	2	2	1

Conclusion

This contribution evaluated the minimum required spatial resolution of early reflections using an image-source model of a shoe-box room and head-tracked binaural Ambisonic playback with the magLS approach. While the resolution of the direct sound was kept at 7^{th} order, the early reflections were played back in orders 0 to 6 and compared to a 7^{th}-order reference. The comparisons were done for speech and noise, 6 or 236 reflections, a source facing towards or away from the listener, as well as in a small and a large room. The minimum required orders were determined as the minimum orders not to yield medians that were significantly different from the reference.

For noise, orders of 6 were required for the small room, while the large room required only 5^{th} order in most cases. There was no effect of the number of reflections. Interestingly, sources facing away from the listener required about 2 orders less than sources facing towards the listener to achieve a median similarity to the reference above 90%. The reduction in sensitivity could be attributed to the increase of the delay between the first two active sound paths when facing away.

For speech, the required orders were generally lower. The most sensitive case was the small room with only 6 reflections with an order of 3. There was no effect of the source orientation, however, the small room required more spatial resolution and the sensitivity decreased with the number of reflections.

In practice, where tens of reflections are typically rendered [19, 20], 2^{nd}-order or even 1^{st}-order resolution might be enough to achieve plausible playback of early reflections in 3DoF applications on headphones.

Further research could investigate similar effects in 6DoF applications, for loudspeaker playback, or using data-based auralization with measured room impulse responses using spatial enhancement algorithms, such as (A)SDM [21, 22, 7, 23] or (HO)SIRR [24, 25]. Moreover, the interaction between different spatial resolutions for direct sound, early reflections, and late reverberation could be worth a look.

Acknowledgments

The authors thank all listeners for their participation in the experiment.

References

[1] S. Werner, F. Klein, A. Neidhardt, U. Sloma, C. Schneiderwind, and K. Brandenburg, "Creation of auditory augmented reality using a position-dynamic binaural synthesis system - technical components, psychoacoustic needs, and perceptual evaluation," *Applied Sciences*, vol. 11, no. 3, p. 1150, 2021.

[2] F. Zotter and M. Frank, *Ambisonics - A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, ser. Springer Topics in Signal Processing. Springer, 2019.

[3] A. Lindau, H.-J. Maempel, and S. Weinzierl, "Minimum brir grid resolution for dynamic binaural synthesis," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3498, 2008.

[4] M. Zaunschirm, M. Frank, and F. Zotter, "Binaural rendering with measured room responses: First-order ambisonic microphone vs. dummy head," *Applied Sciences*, vol. 10, no. 5, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/5/1631

[5] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments," *Acta Acustica united with Acustica*, vol. 98, no. 5, pp. 804–810, 2012.

[6] I. Engel, C. Henry, S. V. Amengual Gari, P. W. Robinson, and L. Picinali, "Perceptual implications of different ambisonics-based methods for binaural reverberation," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 895–910, 2021.

[7] M. Zaunschirm, M. Frank, and F. Zotter, "Brir synthesis using first-order microphone arrays," in *prepr. 9944, 144th AES Conv.*, Milano, 2018.

[8] K. Enge, M. Frank, and R. Höldrich, "Listening experiment on the plausibility of acoustic modeling in virtual reality," 2020.

[9] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses," in *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[10] T. Lübeck, J. M. Arend, and C. Pörschmann, "Binaural reproduction of dummy head and spherical microphone array data - a perceptual study on the minimum required spatial resolution," *The Journal of the Acoustical Society of America*, vol. 151, no. 1, pp. 467–483, 2022.

[11] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[12] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616–3627, 2018.

[13] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Fortschritte der Akustik - DAGA*, Munich, March 2018.

[14] D. Perinovic and M. Frank, "Spatial resolution of diffuse reverberation in binaural ambisonic playback," in *Fortschritte der Akustik - DAGA*, Munich, March 2021.

[15] M. Romanov, P. Berghold, M. Frank, D. Rudrich, M. Zaunschirm, and F. Zotter, "Implementation and Evaluation of a Low-Cost Headtracker for Binaural Synthesis," in *Audio Engineering Society Convention 142*, May 2017.

[16] M. Frank and M. Brandner, "Perceptual Evaluation of Spatial Resolution in Directivity Patterns," in *Fortschritte der Akustik, DAGA*, Rostock, Mar. 2019.

[17] ——, "Perceptual Evaluation of Spatial Resolution in Directivity Patterns 2: coincident source/listener positions," in *International Conference on Spatial Audio, ICSA*, 2019.

[18] EBU, "EBU SQAM CD: Sound Quality Assessment Material recordings for subjective tests," 2008. [Online]. Available: https://tech.ebu.ch/publications/sqamcd

[19] S. Clapp and B. U. Seeber, "Sound Localization in Partially Updated Room Auralizations," in *Fortschritte der Akustik, DAGA*, 2016, pp. 558–560.

[20] M. Frank, D. Rudrich, and M. Brandner, "Augmented Practice-Room - Augmented Acoustics in Music Education," in *Fortschritte der Akustik, DAGA*, Rostock, Mar. 2020.

[21] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial Decomposition Method for Room Impulse Responses," *Journal of the Audio Engineering Society*, vol. 61, no. 1/2, pp. 17–28, January 2013.

[22] M. Frank and F. Zotter, "Spatial impression and directional resolution in the reproduction of reverberation," *Fortschritte der Akustik - DEGA*, 2016.

[23] S. V. Amengual GarÃ, J. M. Arend, P. T. Calamia, and P. W. Robinson, "Optimizations of the spatial decomposition method for binaural reproduction," *J. Audio Eng. Soc*, vol. 68, no. 12, pp. 959–976, 2021. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=21010

[24] V. Pulkki, J. Merimaa, and T. Lokki, "Reproduction of reverberation with spatial impulse response rendering," *Journal of the Audio Engineering Society*, vol. 61, no. 1/2, pp. 17–28, May 2004.

[25] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger, and M. Marschall, "Higher-order spatial impulse response rendering: Investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution," *J. Audio Eng. Soc*, vol. 68, no. 5, pp. 338–354, 2020. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20852