Investigation of the minimum required spatial resolution of moving sound sources

Tim Lübeck[1,2] and Christoph Pörschmann[1]
[1] *Institute of Computer and Communication Technology, TH Köln D-50679 Cologne, Germany*
[2] *Audio Communication Group, TU Berlin, D-10587 Berlin, Germany*
Email: tim.luebeck@th-koeln.de

Introduction

The spatial resolution is an important criterion for the design of virtual acoustic environments (VAEs) for virtual and augmented reality applications. Higher spatial resolution increases the localization of sound sources and improves spatial attributes such as externalization, or envelopment of a VAE. However, higher spatial resolution usually requires more computation and storage. In the case of sound field synthesis based on measurements, it also requires time-consuming recording procedures. Therefore, finding an adequate level of spatial resolution without significantly degrading the perceptual quality of VAEs is an essential research topic.

The spatial resolution has often been investigated in terms of the minimum audible angle (MAA), which describes the smallest detectable angle between two sound incidences [1]. An alternative way to describe the spatial resolution is the order of the spherical harmonics decomposition of a sound field. In particular, in the literature on Ambisonics and spherical array processing, the SH representation of sound fields is a fundamental concept. Each sound field can be represented by frequency-dependent SH coefficients. Thereby, each SH coefficient is associated with one of a set of spherical harmonics which can be considered orthogonal polar patterns. In this context, each spherical harmonic represents one directional component of the sound field. The more spherical harmonics are used to represent the sound field, the more spatial details can be encoded. This is why the SH order can be used to define and scale the spatial resolution of sound fields.

Several studies investigated how the SH order N of binaural renderings from spherical microphone array (SMA) captures influence the quality of VAEs in terms of spaciousness, timbre, and further spatial attributes [2], [3],[4]. It was found that orders of $N = 7-12$ are required to achieve auralization similar to those based on dummy head measurements. Other studies determined thresholds of SH orders N which are required to achieve auralizations that are indistinguishable from a high-resolution reference [5]. It was found that for renderings based on SMA captures, orders of $N = 19 - 20$ are required for indistinguishability to a $N = 29$ rendering. For SH interpolation of dummy head measurements in reverberant sound fields, orders of $N = 11 - 13$ are required. Similar perceptual evaluations were conducted in which the minimum required orders for SH interpolation of head-related transfer functions (HRTFs) were examined [6], [7]. In other studies, the SH order was used as a measure for spatial resolution of source directivities [8, 9]. The authors found, that orders of $N = 3 - 5$ are required to be indistinguishable from a reference at $N = 7$. For the directivity of the human voice, an average order of $N = 8.4$ is required to be indistinguishable from a reference of $N = 35$ [10]. The latter studies all examined non-moving (stationary) sound sources. However, in a real-life scenario, sound sources may also move or rotate around the listener. The influence of spatial resolution of moving sound sources is hence an interesting research question. Several studies already investigated the localization ability of moving sound sources. A comprehensive review of perceptual aspects of sound motion is presented in [11]. The counterpart to the MAA of stationary sound sources is the minimum audible movement angle (MAMA). It has been defined as the minimum angle a sound source needs to be moved to be distinguished from a stationary sound source. It has been shown that the MAMA is smaller than the MAA when measured under comparable conditions [12, 13, 14]. Furthermore, the minimum movement audible angle (MMAA) has been introduced [15]. It is defined as the smallest angular difference between two simultaneously moving sound sources that is required to be able to be perceived as coming from two different directions. It was shown that the MMAA is smaller when the listener moves, compared to the MMAA due to source motion. Hence, the movement of sound sources certainly affects the localization ability, but to the best of our knowledge, no study investigated the influence of spatial resolution on sound motion in form of the SH order.

In this work, we present the results of a listening experiment that determines a minimum required spatial resolution of moving sound sources which is required for indistinguishability to a high spatial resolution reference. The study intends to solely examine the perceptual influence of spatial resolution, without any technical artifacts due to sampling. The limitations in the synthesis of sound fields of moving sound sources that have been captured with SMA captures, for example, were discussed in [16]. We hence defined the spatial resolution as the SH truncation order of the SH coefficients of an HRTF set. A densely sampled HRTF set is decomposed into its SH components up to an order of $N = 35$. At an order of $N = 35$ any sampling or capturing artifacts that would arise in SMA renderings or sparse sampling of HRTFs can be neglected. To scale the spatial resolution of the HRTFs we then truncated the SH order series at orders $N = 1 - 34$. In an ABX listening experiment, participants compared the order-limited HRTFs to the $N = 35$ order reference. By successively increasing the order we determined thresholds (points of subjective equality (PSEs)) where no differences between order-limited and a high-resolution reference is detectable. To investigate how the movement of the source influences the threshold, we conducted the test for stationary sources at $\phi = 90°$ [1] and sound sources moving from $\phi = 135°$ to $\phi = 45°$ around the listener with a fixed distance. Both conditions were tested with a noise, speech, and castanets test signal.

Method

Stimuli

In the listening experiment, we presented stationary sound sources from $\phi = 90°$ and sound sources moving from $\phi = 135°$ to $\phi = 45°$ around the listener with a fixed distance. The binaural signals for each head orientation were pre-calculated in Matlab and played back by the SoundScape Renderer [17, 18] in binaural playback mode, which fades between the signals according to the instantaneous head orientation of the listener. To save working memory we only allowed head orientations between 60° and 300°. Each sound source was synthesized by convolution with the corresponding HRTF of a KU100 dummy head [19] and a snipped of the dry test signals (castanets, speech, and noise). For the stationary sound source, the same HRTF was picked for one head orientation. The set for all head orientations was calculated by inverse rotation.

To synthesize the moving sound source we convolved 90 HRTFs (for each of the angles $\phi = 135°$ to $\phi = 45°$, in 1° steps) with the test signal. To yield a binaural signal of a source that moves with a speed of 2 kmh and a distance of 1.75 m around the listener, we concatenated a signal by successively picking 2638 samples from the corresponding convolved signal. For a smooth and artifact-free rotation, we applied fading with a Hanning window with an overlap of 512 samples. This processing was repeated for each head orientation.

The entire processing was done for 35 different sets of HRTFs: the KU100 HRTFs [19] were measured on a 2702 sampling point Lebedev grid which allows SH decomposition at an order of $N = 44$. We can neglect any artifacts due to sparse sampling and spatial aliasing. We chose a maximum order of $N = 35$ which is enough for the entire audible bandwidth. As described in Sec. 1 we then truncated the $N = 35$ set at the orders $N = 1$ to $N = 34$, resulting in 35 HRTF sets of increasing spatial resolution. Unfortunately, truncating the higher orders also leads to a loss of energy at higher time-frequencies which introduces audible low-pass characteristics. Hence, to exclusively test the influence of the spatial resolution, we applied diffuse field equalization, according to [20] to compensate for the spectral roll-off.

[1] ϕ denotes the horizontal angle ranging from 0° to 359°, and θ the colatitude angle, ranging from 0 to 180

Setup

As described above, in the listening experiment we applied dynamic binaural synthesis using the SoundScape Renderer (v.0.3.4) [17, 18] in binaural playback mode. The listeners' head orientation was tracked with a Polhemus Fastrak at a sampling rate of 120 Hz. The experiment was performed in the anechoic chamber of TH Köln with a background noise level of less than 20 dB(A). As a digital-analog converter, we used an RME Fireface UFX at 48kHz and a buffer size of 256 samples and Sennheiser HD600 headphones for playback with a playback level of about 62dB(A). We equalized the binaural chain of the Neumann KU100 dummy head and Sennheiser HD600 headphones using a 2048 tap minimum phase compensation filter designed according to a regularization method proposed in [21].

Test design

To determine thresholds of indistinguishability as the PSE, we employed the ABX three-interval/two-alternative forced-choice (3I/AFC) test design [22, 23] together with an adaptive one-up one-down staircase procedure [24]. The participants were provided with a tablet computer displaying a graphical user interface with 3 buttons labeled A, X, and B. In each trial, three sequences were played back to the participant, two of which were the same. Either the reference stimulus or the low-spatial-resolution stimulus was assigned randomly to the middle interval X, such that either A or B was the same stimulus as X. This ensures that the order-limited stimulus or the reference was played back two times and that participants could directly compare between both conditions. After listening to the sequence, the participants were asked to choose whether A or B equals X. According to the one-one down staircase method, if the participants were correct and could indicate the difference between the intervals, the stimulus based on the HRTF with the next higher spatial order was assigned in the next trial. If they gave a wrong answer, i.e., they could not indicate any differences, the stimuli based on the next lower order N were picked for the next trial. Each run started with the low-resolution stimuli of order $N = 1$ and was terminated after 12 reversals. One reversal is defined as a correct decision followed by a wrong decision or vice versa. Each ABX sequence was automatically played back one time and, during the playback, the participants were free to move their heads if it helped them to distinguish between the auralization.

Participants

11 participants (1 female, 10 male) with a mean age of 33.2 years took part in the experiment. All of them were experienced listeners, and most of them were staff from the laboratory at TH Köln.

Data Analysis

We determined the PSEs by averaging the last 9 reversals and thus omitted the first 3 reversals. In the following, we refer to discrete values of the PSEs, because the signal processing requires integer values of N. A Shaprio-Wilk test for normality distribution failed to reject for all conditions. Therefore, we analyzed the data with a within-subject analysis of variance (ANOVA) with the between-subject factors test signal (noise, speech, and castanets) and source kind (moving, stationary).

Results

A graphical overview of the results is presented as boxplots in Fig. 1. It can be observed that for all test signals, the median values for the moving source PSEs are notably smaller than for the stationary sources. Whereas the median values for the speech and castanets are in the same range, the PSEs for the noise signal are about 6 orders higher.

The mean values are depicted in Fig. 2. The same observation as from the boxplots can be made. For all test signals the mean values are notably smaller for the moving sound sources compared to the stationary sound sources. All mean values are summarized in Tab. 1. The results of the ANOVA support these observations. The source kind (moving vs. stationary) was revealed as a significant effect $F(1, 10) = 44.27$, $p < .001$, $\eta_p^2 = .82$, $\epsilon = 1.0$. Hence, moving sources require significantly fewer SH orders than stationary sources. Furthermore, the ANOVA revealed the test signal as a significant effect $F(2, 20) = 25.22$, $p < .001$, $\eta_p^2 = .72$, $\epsilon = .82$.

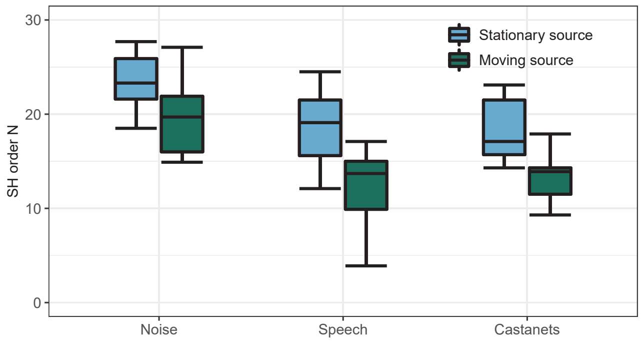

Figure 1: Boxplots of the inter-individual variation in the PSEs for each test signal separately. The boxes of blue color (left) depict the results for the stationary sound source, the green boxes (right) the results for the moving sources. The medians are indicated as a black line.

Table 1: Mean values of the PSEs across subjects and corresponding 95% confidence intervals.

Signal	Stationary	Moving
Noise	23.5 ±2.03	19.57 ±2.66
Speech	18.31 ±2.68	12.43 ±2.62
Castanets	18.54 ±2.17	13.23 ±1.69

Discussion and Conclusion

The results show that the auralization of moving sound sources requires significantly fewer SH orders than non-moving sound sources to be indistinguishable from the corresponding $N = 35$ auralization. This indicates that the auditory spatial perception is more sensitive to non-

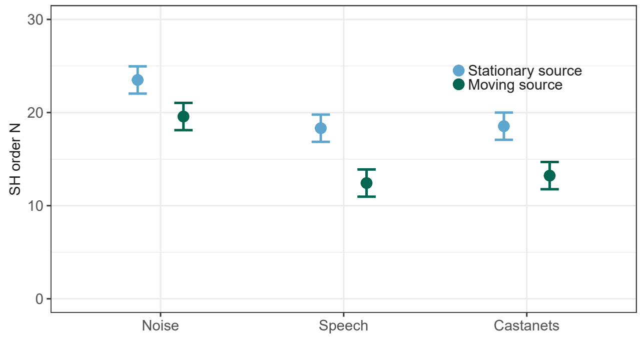

Figure 2: Mean values of the PSEs across subjects and the corresponding 95% confidence intervals. The blue color indicates the results for the stationary sound source, the green color the results for the moving sources

moving sound sources than to moving sound sources. This is in line with the findings that the MAMA is smaller than the MAA [12, 13, 14]. This could be explained by the fact that the human auditory system is relatively sensitive to interaural temporal differences but rather insensitive to the dynamic variation of these differences, see [25]. Another thing that may have influenced the results of the study is that the trajectory of the moving source always started at $\phi = 135°$ and ends at $\phi = 45°$. As a result, participants always directly compared the endpoints of the first stimulus with the start point of the next stimulus, which was not the case for the stationary sources. To investigate this, a follow-up study conducting similar listening experiments with varying trajectories is suggested. In addition, the present study only evaluates sources moving with the same velocity. Past studies have shown that the perception of motion is highly dependent on the velocity of motion and the stimulus duration. Evaluating different velocities is therefore a topic for future research. Moreover, we neglect the influence of distance variation. It might be interesting to evaluate sources that not only rotate at a fixed distance around the listener but also move with varying distances. Effects such as the Doppler shift or sound level variations could further influence motion perception. Finally, testing sources at different heights is left for future research. Past studies have shown that the MAMA is larger for sources moving along vertical trajectories, compared to sources moving in the horizontal plane [12]. Thus, certainly, the perception of motion is dependent on the exact trajectory. However, since the auditory system seems to be more sensitive to horizontal trajectories, the thresholds found in this study can be considered rather conservative.

References

[1] A. W. Mills, "On the Minimum Audible Angle," *Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, 1958.

[2] B. Bernschütz, "Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording," p. 264, 2016.

[3] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, “Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution,” *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2711–2721, 2013.

[4] J. Ahrens and C. Andersson, “Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre,” *The Journal of the Acoustical Society of America*, vol. 145, no. April, pp. 2783–2794, 2019.

[5] 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, 2021.

[6] J. M. Arend, F. Brinkmann, and C. Pörschmann, “Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions,” *Journal of the Audio Engineering Society*, vol. 69, no. 1/2, pp. 104–117, 2021.

[7] C. W. Pike, “Evaluating the Perceived Quality of Binaural Technology,” Ph.D. dissertation, University of York, 2019.

[8] M. Frank and M. Brandner, “Perceptual Evaluation of Spatial Resolution in Directivity Patterns,” in *Proceedings of the 45th DAGA*, no. April 2019, Rostock, GE, 2019, pp. 74–77.

[9] M. Frank and M. Brandne, “Perceptual Evaluation of Spatial Resolution in Directivity Patterns 2: coincident source/listener positions,” in *5th International Conference on Spatial Audio ICSA*, no. September, 2019, pp. 1–8.

[10] A. Quélennec and P. Luizard, “Pilot study on the influence of spatial resolution of human voice directivity on speech perception,” *Acta Acustica*, vol. 6, 2022.

[11] S. Carlile and J. Leung, “The Perception of Auditory Motion,” *Trends in Hearing*, vol. 20, pp. 1–19, 2016.

[12] D. W. Grantham, B. W. Y. Hornsby, and E. A. Erpenbeck, “Auditory spatial resolution in horizontal, vertical, and diagonal planes,” *The Journal of the Acoustical Society of America*, vol. 114, no. 2, pp. 1009–1022, 2003.

[13] D. W. Grantham, “Detection and discrimination of simulated motion of auditory targets in the horizontal planea),” *Journal of the Acoustical Society of America*, vol. 79, no. 6, pp. 1939–1949, 1986.

[14] K. Saberi and D. R. Perrott, “Minimum audible movement angles as a function of sound source trajectory,” *The Journal of the Acoustical Society of America*, vol. 88, no. 6, pp. 2639–2644, 1990.

[15] W. O. Brimijoin and M. A. Akeroyd, “The moving minimum audible angle is smaller during self motion than during source motion,” *Frontiers in Neuroscience*, vol. 8, no. SEP, pp. 1–8, 2014.

[16] J. Ahrens and S. Spors, “Artifacts in the sound field of a moving sound source reconstructed from a microphone array recording,” in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, Ed., New Paltz, NY, 2009, pp. 165–168.

[17] M. Geier, J. Ahrens, and S. Spors, “The SoundScape Renderer,” 2019. [Online]. Available: http://spatialaudio.net/ssr/

[18] M. Geier and S. Spors, “Spatial Audio with the SoundScape Renderer,” *27th Tonmeistertagung - VDT International Convention*, p. 2012, 2012.

[19] B. Bernschütz, “A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100,” in *Proceedings of the 39th DAGA*, Meran, 2013, pp. 592—-595. [Online]. Available: http://www.audiogroup.web.fh-koeln.de/FILES/AIA-DAGA2013{_}HRIRs.pdf

[20] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, “Spectral equalization in binaural signals represented by order-truncated spherical harmonics,” *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, 2017. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.4983652

[21] V. Erbes, H. Wierstorf, M. Geier, and S. Spors, “Free database of low-frequency corrected head-related transfer functions and headphone compensation filter,” in *Proceedings of 142nd AES Convention*, 2017, pp. 1–5.

[22] F. A. Kingdom and N. Prins, *Psychophysics: A practical Introduction*, 1st ed. London, UK: Academic Press is an imprint of Elsevier, 2010.

[23] H. Levitt, “Transformed Up-Down Methods in Psychoacoustics,” *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 467–477, 1971.

[24] T. Meese, “Using the standard staircase to measure the point of subjective equality: A guide based on computer simulations,” *Perception & Psychophysics*, vol. 25, no. 1, pp. 16–18, 1995.

[25] D. W. Grantham and F. L. Wightman, “Detectability of varying interaural temporal differencesa),” *Journal of the Acoustical Society of America*, vol. 63, no. 2, pp. 511–523, 1978.