Evaluation of Real-time Implementation of 3D Multichannel Audio Rendering Methods

Merle Gerken[1], Giso Grimm[1], Volker Hohmann[1]

[1] *Auditory Signal Processing and Cluster of Excellence "Hearing4all", Department of Medical Physics and Acoustics, University of Oldenburg, Germany. E-mail: merle.gerken@uol.de, g.grimm@uol.de, volker.hohmann@uol.de*

Introduction

Virtual acoustics are increasingly applied in hearing aid research to achieve a reproducible and immersive evaluation of hearing devices. For this it is of great importance to have three-dimensional (3D) methods available, because the additional elevation is important to create immersion. Therefore, real-time implementations of 3D rendering methods such as Vector Base Amplitude Panning (VBAP) and Higher Order Ambisonics (HOA) are implemented in the Toolbox for Acoustic Scene Creation and Rendering (TASCAR) [1]. The perceptual and physical limitations of 2D multi-channel reproduction has been assessed [2, 3]. Little research can be found regarding the perceptual evaluation of 3D reproduction methods, and most studies either focus on 3D microphone techniques [4], technical analysis [5] or mixed order systems [6].

This study aims at evaluating the perceptual performance of the implementation of 3D rendering methods in an interactive rendering system for moving sources and receivers, given a reproduction system with only sparse distribution of loudspeakers. The performance is assessed in terms of the absolute localization of sound sources and the perceptive spatial resolution.

In this paper, first the implemented spatial reproduction methods and the experimental paradigm is described. Results of localization and spatial resolution are shown and discussed.

Methods

Rendering methods

Different rendering methods were available respectively newly implemented in TASCAR and evaluated in this study. These include 2D methods for reproduction via a horizontal loudspeaker array, and 3D methods for reproduction via loudspeakers spaced irregularly on a sphere.

Nearest speaker selection

The nearest speaker selection (NSP) method, which was available in TASCAR, maps a given virtual sound source to the loudspeaker that has the smallest angular distance, so one channel is used to play back the sound [1].

Vector base amplitude panning 3D

Vector base amplitude panning in three dimensions (VBAP 3D) according to [7] was additionally implemented in TASCAR. A phantom sound source is created by applying different gains on three channels around the virtual source. Based on the positions of the loudspeakers, a convex hull is calculated, which consists of multiple triangles where the corners are defined by the loudspeaker positions. For the creation of a virtual sound source, the closest triangle is selected, and the gains are calculated based on the virtual source vector and the loudspeaker vectors. The gains are then scaled using the sound power C [7].

Higher order Ambisonics 2D

This method, which was available in TASCAR, uses Higher Order Ambisonics [8] to render virtual sources on the horizontal plane. It is available either with basic decoding (HOA 2D basic) or with max $\mathbf{r}_E$ decoding (HOA 2D max $\mathbf{r}_E$), where the vector of the maximum energy points towards the virtual source position. In this study, 7th order Ambisonics was used.

Higher order Ambisonics 3D

The HOA decoder options were extended by a 3D version for arbitrary 3D speaker layouts. Decoding is available either with the Ambisonics mode matching method using the Moore-Penrose pseudoinverse (HOA 3D max $\mathbf{r}_E$ pinv), or the All round Ambisonic decoding (AllRAD) method via regular virtual speakers rendered with VBAP (HOA 3D max $\mathbf{r}_E$ AllRAD) [9]. For the evaluation in this study, both versions were combined with max $\mathbf{r}_E$ decoding. Again, 7th order Ambisonics was used, resulting in a slightly under-determined system for elevated sources. The HOA 3D decoders were checked against a reference implementation of the Ambisonics Decoder Toolbox [10].

Apparatus

The evaluated setup consisted of 29 loudspeakers of type Genelec 8020 in four rings of different elevation. The main ring (0° elevation) consisted of 16 loudspeakers, placed every 22.5° starting at 11.25°. In the lower ring of six channels (-40° elevation) the speakers were placed every 60°, starting at 30° azimuth. In the higher ring of six channels (30° elevation) the speakers were also placed every 60°, but started at 0°. One loudspeaker was placed in the center at the top. In addition to the acoustic reproduction, three video projectors were used for the reproduction of a visual pointer (field of view was 300° in azimuth and 2m height). Head movements were tracked using a Qualisys infrared marker tracking system with six cameras for tracking of a crown, which was worn by the test participants. For user input, a Behringer XTouch One MIDI controller was used. The sounds were rendered with TASCAR, and the game engine Blender was used to create the visual pointer. Data logging was also realized in TASCAR. The experimental control was implemented in MATLAB. The participants were sitting in one of two possible sitting positions: One was located in the center, and the other was at an off-central position, on average located 10.5 cm behind, 84.3 cm to the right and 31.0 cm downwards relative to the center position.

Instrumental performance measure

The localization errors evolving from the application of the HOA 3D methods were simulated with the Ambisonics Decoder Toolbox [10] to serve as a reference for the subjective experiments.

Subjective performance measures

Absolute sound source localization

The subjective localization of virtual sound sources was evaluated at the center and off-center listening position. An International Speech Test Signal (ISTS) [11] served as a stimulus, which was rendered to 17 different virtual sound sources (see Figure 1), including nine positions on the height of the main loudspeaker ring (0° elevation), three positions above and five positions below. The participants' head direction was displayed on a screen by projectors as an optical pointer. The subjects were asked to move this pointer to the position where the sound was perceived. After finding this position with the pointer, the subjects confirmed their choice using a push button and the next stimulus became active. In a randomized order, all 17 virtual sources were rendered by the methods HOA 2D max $\mathbf{r}_E$, NSP, VBAP 3D, HOA 3D max $\mathbf{r}_E$ pinv and HOA 3D max $\mathbf{r}_E$ AllRAD.

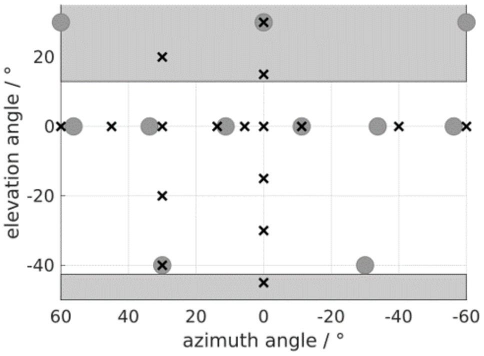

Figure 1: Virtual rendering positions in the absolute localization experiment. The white area indicates the projection area, the grey dots represent the real loudspeaker positions and the black crosses are placed at the virtual sound sources.

Spatial resolution

The perceptive spatial resolution that can be achieved by the rendering methods was measured in terms of the minimum audible angle (MAA) on the azimuth plane. An alternative forced choice experiment with two intervals (2-AFC) was applied. The subjects were sitting in the center position of the lab, and a pink noise served as the stimulus. This stimulus was rendered at different elevation angles by five different methods as shown in Table 1 for two azimuth angles in each trial, and the subjects were asked to respond in which azimuth direction the stimulus was moving using a graphical user interface. The absolute azimuth was randomized between trials in the range of -22.5 to 11.25° to avoid a bias caused by coloration differences. Such coloration differences could potentially result in an underestimation of the MAA for positions very close to physical loudspeaker positions.

Table 1: Measured positions and rendering methods in the minimum audible angle experiment.

Rendering method	Position (azimuth, elevation) / degree
HOA 2D max $\mathbf{r}_E$	(0,0)
HOA 2D basic	(0,0)
HOA 3D max $\mathbf{r}_E$ pinv	(0,0); (0,-30)
HOA 3D max $\mathbf{r}_E$ AllRAD	(0,0); (0,-30)
VBAP 3D	(0,0); (0,-30)

Participants

Ten self-reported normal hearing listeners at the age of 19-26 years participated in the experiments.

Results

Instrumental measures

The angular error based on the energy vector for HOA 3D max $\mathbf{r}_E$ pinv respectively HOA 3D max $\mathbf{r}_E$ AllRAD is shown in Figure 2 and Figure 3, where the highlighted part visualizes the range of measured elevation angles in the subjective experiment. The simulations were conducted with a 7th order Ambisonics system with the same layout as the physical setup. From these results, it can be predicted that the localization error is small on the 0° elevation plane, and larger for elevated virtual positions for both methods. The predicted angular error at elevated sources is slightly larger for the pinv method (approx. 10°) than for the AllRAD decoder (5 to 10°).

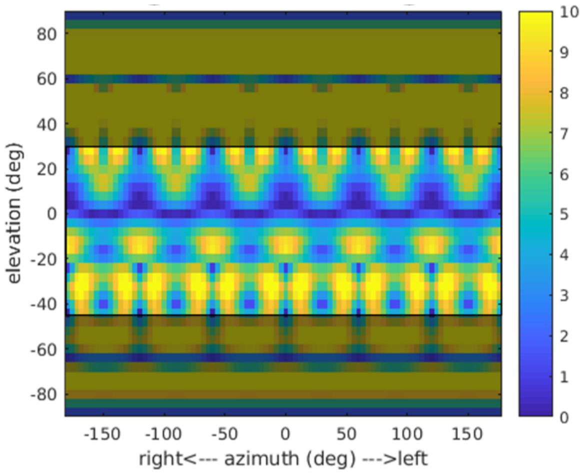

Figure 2: Localization error in degrees in terms of the energy vector for the method HOA 3D max $\mathbf{r}_E$ pinv, obtained with the Ambisonics Decoder Toolbox.

Subjective measures

Absolute sound source localization

Figures 4 - 7 show the results of the absolute localization experiments in terms of azimuth and elevation angle for the center (Figures 4 and 5) and off-center (Figures 6 and 7) listening position, respectively. The error measure on the y-axis is the difference between the virtual source position and the perceived position. The thick line represents the median value, and the bar shows the interquartile range across all participants and source positions. The x-axis is grouped depending on the elevation

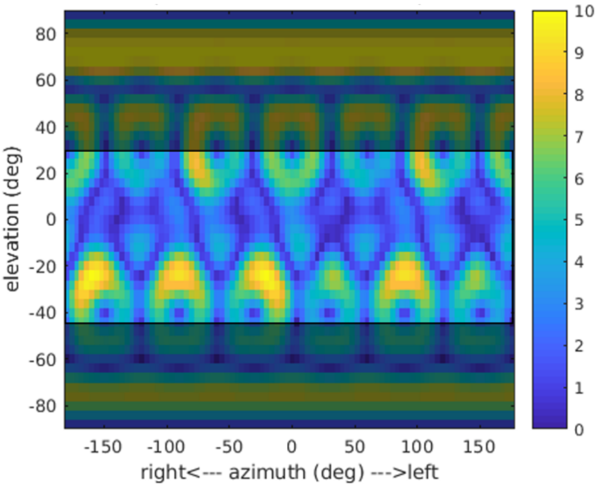

Figure 3: Localization error in degrees in terms of the energy vector for the method HOA 3D max $\mathbf{r}_E$ AllRAD, obtained with the Ambisonics Decoder Toolbox.

angle of the virtual sound sources. The group *elevation 0°* consists of 90 data points per rendering method, *elevation > 0°* consists of 30 data points and *elevation < 0°* consists of 50 points. The plots displaying the elevation error also contain a black line for elevated sources, which is the negative median elevation of all target positions in this group, i.e. the maximum expected error for 2D reproduction.

The azimuthal localization error in the center listening position (Figure 4) is below 5° for all reproduction methods except for the NSP and sources with negative elevation. For virtual sources on the main ring (0° elevation) the error is largest for the NSP rendering method, followed by HOA 3D with AllRAD decoder. The elevation error in the center listening position (Figure 5) is negligible for virtual sources from the main ring. For elevated sources, the error is largest for the 2D reproduction method. Here, the error is very close to the median elevation of the sources represented by the black line, which would be the expected localization error given the small error for virtual sources on the ring. No clear differences can be seen between the other rendering methods, except for NSP in case of sources below the main ring, which results in smallest elevation errors of approximately 10°.

Comparing the results of the off-center position to the results obtained in central listening position, the performance of HOA and VBAP is partially lowered, because these methods work best in the center ("sweet spot"). In contrast, the performance of NSP remains similar. Moreover, the performance of VBAP 3D tends to be better than HOA 3D in this absolute localization task.

Spatial resolution

The subjective spatial resolution results are depicted in Figure 8. The y-axis shows the MAA in degrees as the median and interquartile range over 10 data points per bar. The x-axis is grouped by elevation. The MAA is small for all reproduction methods, and for sources on the main ring approx. 2 to 4°. It can be seen that the MAA is higher for down-shifted sources compared to 0° elevation, and that HOA leads to a substantially higher spatial resolution than VBAP.

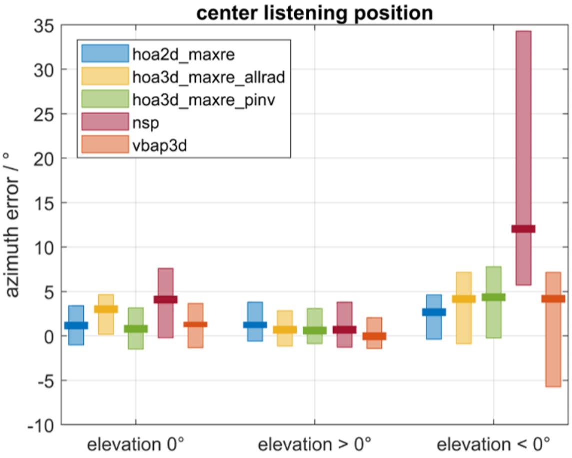

Figure 4: Localization error in azimuth direction for the center listening position.

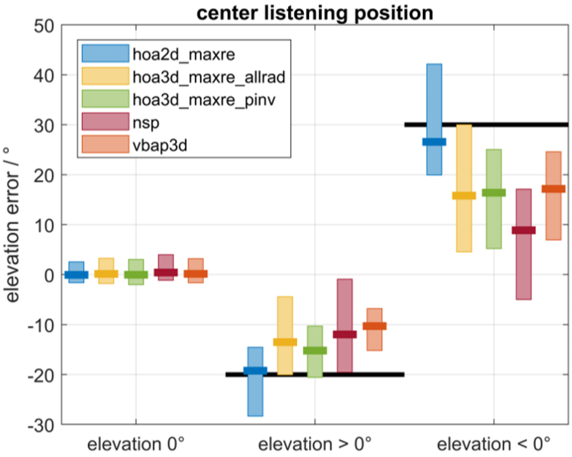

Figure 5: Localization error in elevation direction for the center listening position.

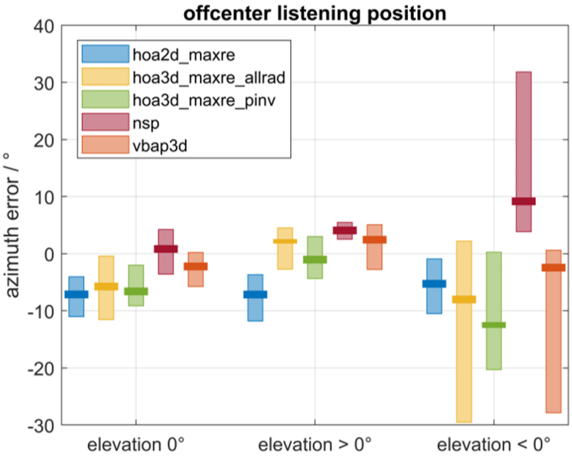

Figure 6: Localization error in azimuth direction for the off-center listening position.

Discussion and conclusions

Some participants reported that they were influenced by the fact that some of the loudspeakers were visible

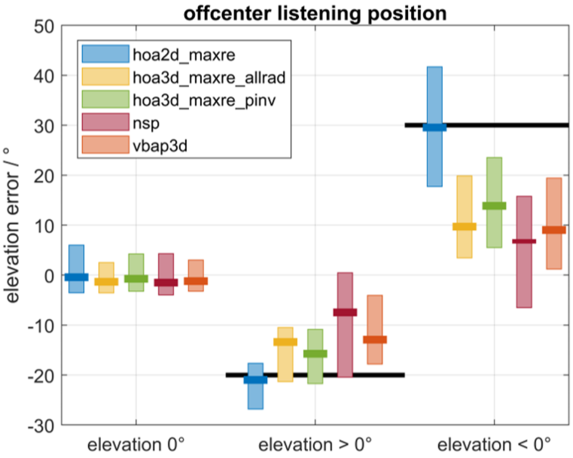

Figure 7: Localization error in elevation direction for the off-center listening position.

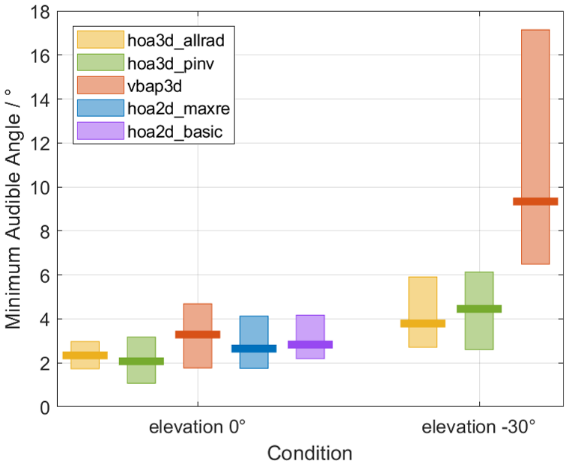

Figure 8: Minimum audible angle in degrees.

through the projection screen. This might lead to a bias in the answers towards the real loudspeaker positions. To lower the impact of the number and position of real loudspeakers on the outcome of the study, the experiments could be repeated in a different loudspeaker setup.
In the localization task the participants were instructed to turn their head towards the virtual sound source. This way the source was always in the direction of highest ITD sensitivity. This was a design decision, to achieve constant human resolution for all tested virtual positions.

In this study, instrumental and subjective methods were applied to evaluate real-time implementations of 2D and 3D audio rendering methods. Both in instrumental and subjective experiments, it was observed that the localization error measures perform best at 0° elevation, which is at the main loudspeaker ring in the tested setup. At this position, the newly implemented 3D methods perform similarly to the 2D methods. Furthermore, the 3D decoders lead to the trend of decreased errors for elevated sources, compared to 2D methods. The observed MAA shows a similar magnitude like with headphone measurements. The absolute localization performance is better when applying VBAP, whereas the subjective spatial resolution is highest for HOA.

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 352015383 – SFB 1330 B1.

References

[1] Grimm, G., Luberadzka, J., & Hohmann, V. (2019). A toolbox for rendering virtual acoustic environments in the context of audiology. Acta Acustica United with Acustica, 105(3), 566-578.

[2] Bertet, S., Daniel, J., Parizet, E., & Warusfel, O. (2013). Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. Acta Acustica United with Acustica, 99(4), 642–657.

[3] Wierstorf, H., Raake, A., & Spors, S. (2017). Assessing localization accuracy in sound field synthesis. The Journal of the Acoustical Society of America, 141(2), 1111–1119. https://doi.org/10.1121/1.4976061

[4] Bertet, S., Daniel, J., Gros, L., Parizet, E., & Warusfel, O. (2007). Investigation of the perceived spatial resolution of higher order Ambisonics sound fields: A subjective evaluation involving virtual and real 3D microphones. Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments.

[5] Samarasinghe, P. N., Poletti, M. A., Salehin, S. M. A., Abhayapala, T. D., & Fazi, F. M. (2013). 3D soundfield reproduction using higher order loudspeakers. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, (1), 306–310. https://doi.org/10.1109/ICASSP.2013.6637658

[6] Favrot, S., Marschall, M., Käsbach, J., Buchholz, J., & Weller, T. (2011). Mixed-order Ambisonics recording and playback for improving horizontal directionality. Audio Engineering Society Convention 131.

[7] Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. Journal of the audio engineering society, 45(6), 456-466.

[8] Daniel, J. (2001). Representation de champs acoustiques, applications ala transmission et ala reproduction de scenes sonores complexes dans un contexte multimedia. Paris: Université Pierre et Marie Curie (Paris VI).

[9] Zotter, F., & Frank, M. (2012). All-round ambisonic panning and decoding. Journal of the audio engineering society, 60(10), 807-820.

[10] Heller, A. J., & Benjamin, E. M. (2014, May). The ambisonic decoder toolbox: Extensions for partial-coverage loudspeaker arrays. In Linux Audio Conference.

[11] Holube, I., Fredelake, S., Vlaming, M., & Kollmeier, B. (2010). Development and analysis of an international speech test signal (ISTS). International journal of audiology, 49(12), 891-903.