Perceptive Evaluation of Sound Field Rotation Methods in the Context of Dynamic Binaural Rendering of Ambisonics Signals

Jorgos Estrella Cazuriaga[1], Jan Plogsties[1], Maximilian Neumayer[2] and Jan Bruemmerstedt[1]
[1] *Fraunhofer IIS, Erlangen*
[2] *TU Berlin*

Introduction

Auditory virtual environments have been a topic for research and development since decades. In recent years, consumer grade hardware and new applications for virtual, augmented and mixed reality have created a large interest in this field. Two essential components of virtual auditory displays are binaural rendering and the ability of audio signal processing to react to changes in the orientation of the user's head. This so-called "head-tracking" is enabled by a combination of sensors that provide information about the rotational movement. Most head-tracking devices provide data for three-degrees-of-freedom (3DOF), namely rotational movement around three axes, i.e. yaw, pitch and roll [1, 2]. Translational movement along the three room axes can be achieved using additional transmitters and sensors resulting in 6DOF.
The scope of this work is limited to head rotation and the quality aspects of different implementations.
Audio signals to be presented in auditory virtual displays can have different formats. For conventional music and film sound mixes, channel-based formats are most commonly used, e.g. 5.1 surround sound. More recently, object based mixes have become more important in cinema and gaming applications. Additionally, Ambisonics as a sound field representation has received more attention in virtual reality applications in the last years. It captures the sound field at one point in space, using a spherical harmonics representation. In order to decode an Ambisonics signal, each signal component needs to be fed into a linear gain matrix. The resulting signals can be played back via an arbitrary loudspeaker setup [3].
Binaural rendering is based on the concept of head-related transfer functions (HRTFs). HRTFs are measured on human or artificial heads. They capture the free-field transfer characteristics of the acoustic path from a point in space to the ear canal. HRTFs are typically represented as pairs of impulse responses for the left and right ear from several directions stored in databases. Rendering can be done in a straightforward matter, using convolution of the audio input signals with pairs of HRTFs. The resulting left and right signals are played back over headphones and aim to create a signal that can be localized by the listener from the intended direction [4].
Dynamic binaural rendering requires updating the audio signal dependent on the listener's orientation. There are different methods how this can be done with specific advantages and drawbacks.
The goal of this work is to study perceptual quality of different rotation methods. The specific research question is: What is the optimal way to rotate Ambisonics audio material dependent on the user's head orientation for binaural reproduction over headphones?
In this paper an experimental design is presented to compare different rendering methods with respect to the their behaviour on head-tracking. The results are presented and analyzed, followed by discussion and conclusion.

Implementation

The user interface and audio processing are developed using MAX MSP (see Figure 2).
The input stimuli are stored as 16 channel Ambisonics files, also known as third order Ambisonics. The output signal is a binaural signal presented over headphones.
The signal flow is designed as a single chain while the modules rendering different soundfield rotation approaches are alternatively activated according to the user's current selection. Figure 1 shows this signal processing chain.
The virtual loudspeaker configuration is chosen to contain eight loudspeakers uniformly spaced around the listener in the horizontal plane, as well as four loudspeakers each in the planes elevated $\pm 45°$ from the listener.
Rotation can be applied to the signal in three different modules:

- The Ambisonics Rotator VST plug-in rotates the soundfield by applying a matrix multiplication to the Ambisonics channels. The required rotation matrices are computed using an efficient algorithm by Ivanic and Ruedenberg [5], which constructs the higher order rotation matrix out of the elements of the first order 3×3 rotation matrix.

- The VBAP Rotator receives loudspeaker feeds and rotates the soundfield by creating phantom sources within the fixed loudspeaker layout. It uses the triangulation and panning of each virtual speaker according to [6].

- The Binauralizer receives the loudspeaker feeds and processes them to obtain two headphone signals. If rotation is enabled it receives a matching set of HRTFs from the database with respect to the headtracker data. If rotation is disabled it will use fixed HRTFs with respect to the fixed positions of the virtual loudspeakers. The system is implemented by using the SPAT software [7] with

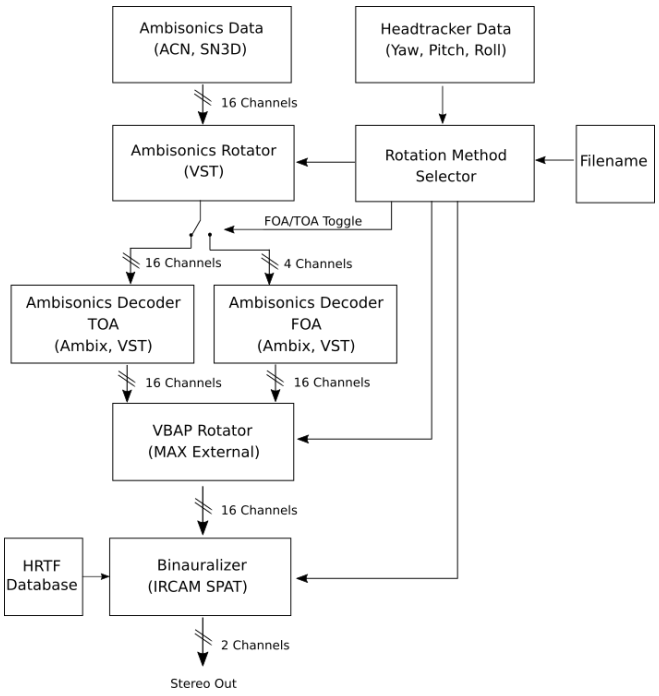

Figure 1: Audio signal flowchart. Note that only the rotation data is toggled.

HRTFs recorded with a KEMAR mannequin from the CIPIC Database [8].

Listening Test

Stimuli

The set of audio stimuli contained seven different excerpts covering a wide spectrum of different sounds. It includes spatial real world recordings as well as synthetically generated signals. The two synthetic audio stimuli are pink noise and white noise bursts and were panned using the ambiX - Ambisonic plug-in suite [9] which allows the positioning of mono sources at an arbitrary position on a sphere and encoding to third order Ambisonics. The two musical stimuli were mixed by sound engineers. They contain a guitar sound and a musical mix. Two real world recordings are used, captured with an Eigenmike soundfield microphone containing a yelling stadium crowd and a scene were a helicopter flies above the listener. One stimulus contains a male speaker reading a text in English. The audio items have a length between fourteen and twenty seconds and were encoded as third order Ambisonics in the Ambix format [10] with a sampling frequency of 44.1 kHz.

Materials and Apparatus

The test software is running on a HP Elitebook notebook with 16GB of RAM and a Intel Core i7-6500 dual-core 2.5 GHz processor. The buffer size of the audio processing is set to 1024 samples with a sampling frequency of 44.1 kHz. A RME Madiface Pro serves as audio interface which is connected to a STAX Studio Monitor amplifier with diffuse field compensation switched on. STAX Lambda Pro New headphones are used.

The participants are free to adjust the loudness of the headphone signal.

Participants

Twenty participants (2 female, 18 male) volunteered to participate in the experiment. The participants were aged between 20 and 52 years (M = 32.60, SD = 8.60). Nineteen of them reported that they already conducted at least one listening test before. Ten participants reported that they already participated in a dynamic binaural listening test before and ten out of twenty stated out that they are specialists in spatial audio. Seven reported to be experts in judging timbre.

Procedure

The listening test session took part in a small office room. The participants were sitting in front of a computer monitor showing the experiment software (see Figure 2). First, they were asked to fill out a short questionnaire to retrieve information about their gender, age, if they have conducted a listening test before, and a self-rating of their listening test expertise.Then, they were informed that this is an experiment about spatial audio and that a headtracker is attached to their headphones. They were also told that they can rotate their head freely and that the soundfield should stay stable when they move their head. The experimenter took care that the participants wore the headphones with the correct orientation and left the room. From now on all instructions were presented only by the experiment software and all interaction with the test only happend inside the experiment software.
The participants were asked to rate their subjective impression of defects included in the sound stimulus presented. For this purpose they were able to switch quickly between the different roation methods while the stimulus was played. The participants gave their ratings by using sliders wich were denoted with a scale from 0 to 100 points, where 100 to 80 points were labelled as "no perceptible defects", 80 to 60 points as "some perceptible but not annoying defects", 60 to 40 points as "some slightly annoying defects", 40 to 20 points as "a number of annoying defects" and 20 to 0 points as "too many annoying defects".
The order of appearance of the stimuli, as well as the assignment of different rotation methods to the sliders were randomised.

Results and Statistics

Twenty participants took part in the listening test. The ratings from one participant had to be discarded because the participant stated that he was over challenged by the listening test. Hence, the results of 19 participants, each one listening to seven audio items with four different rotation methods, respectively, were evaluated. This leads to a sum of 133 observations for each rotation method and 532 observations in total.
A Welch's t-test was chosen, to check among which groups a statistically significant difference exists (see Table 1). Before conducting this test, a Bonferroni correc-

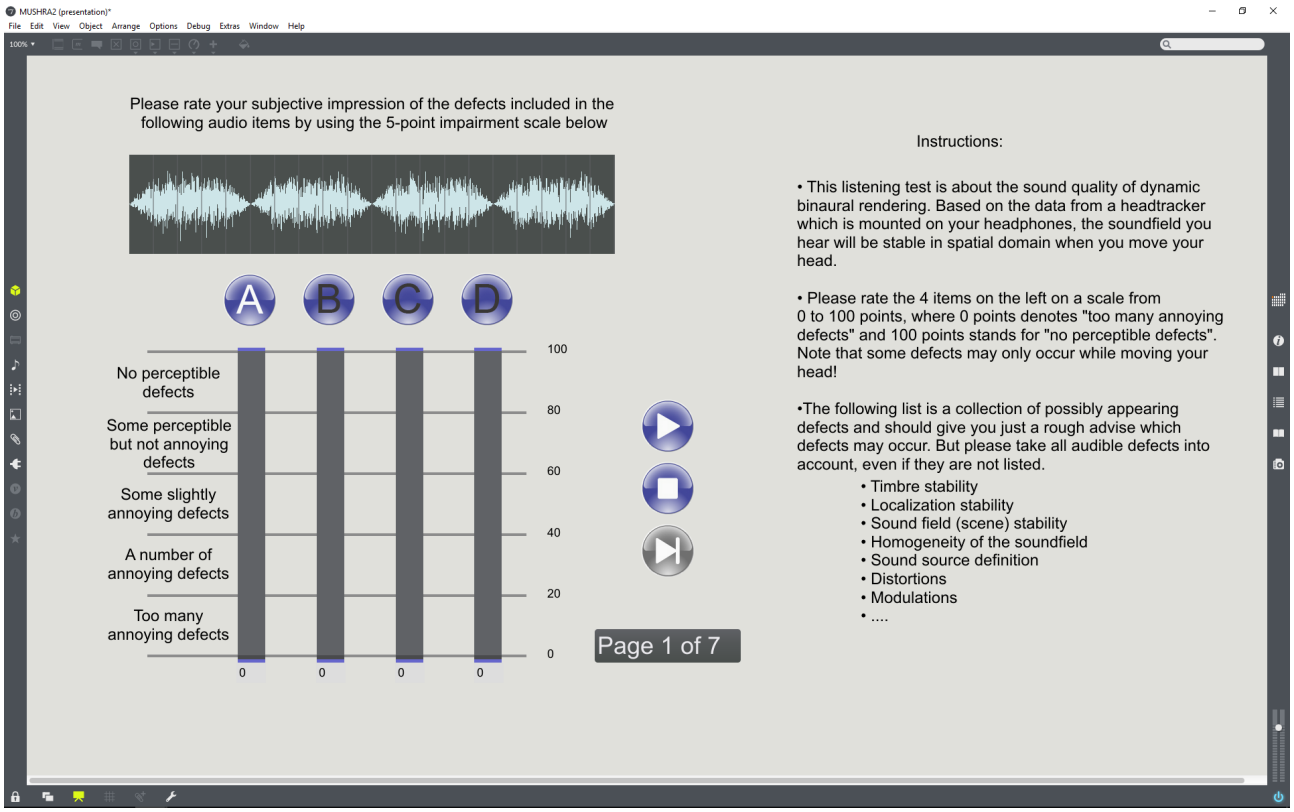

Figure 2: Listening test user interface.

tion had to be applied to the significance level to counteract the problem of multiple comparisons. Six observations are made in total, lowering the significance level of α which was set to 0.05 to a new value of $\alpha = 0.0083$. The t-test showed a significant difference in the ratings for the TOA rotation method ($M = 78.22, SD = 15.56$) and the rotation method, which replaces HRTFs ($M = 58.95, SD = 22.74$). There was also a significant difference in the ratings for the FOA rotation method ($M = 69.15, SD = 20.38$) and the HRTF rotation method. Comparing the ratings for the VBAP rotation method ($M = 73.95, SD = 20.00$) with the HRTF rotation method showed also a significant difference. No significant difference was found comparing the FOA with the VBAP rotation method, as well as the comparison of the TOA with the VBAP rotation method. Furthermore a significant difference was found between the FOA and TOA rotation method.

To assess the strength of the effects Cohen's d was calculated. Cohen's d is a measurement expressing score distances in units of variability [11]. According to S. Sawilowsky [12] a Cohen's d of 0.2 is considered a small effect, a value of 0.5 a medium effect, 0.8 as a large effect, and 1.2 a very large effect. Cohen's d indicates a large effect (d = 1) for the comparison of the TOA rotation method with the HRTF rotation method. Medium effects are observed between the two Ambisonics rotation methods (d = 0.5), between the FOA and HRTF rotation method (d = 0.5), as well as the observation of the VBAP with the HRTF rotation method (d = 0.7).

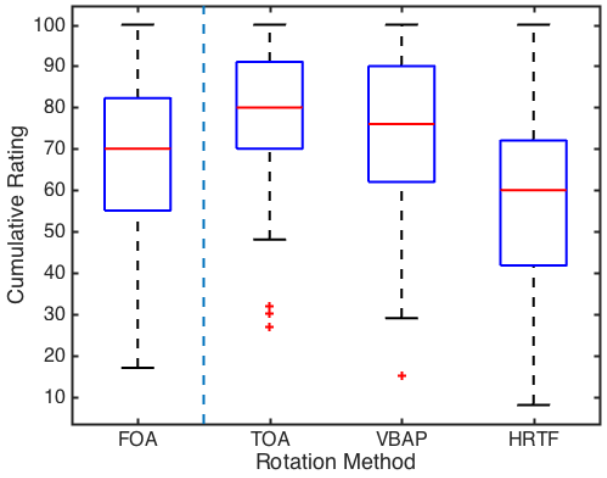

Figure 3: Boxplot of the test results. The y-axis marks the points on the rating scale of the listening test.

Discussion and Conclusion

In this work, different methods of sound field rotation when using binaural rendering were implemented and evaluated with respect to their perceptual quality. A listening test was designed and conducted with 20 participants. It was found that listeners are able to discriminate between the different rotation methods. Rotation in Ambisonics domain and using VBAP received the highest ratings in overall quality. Exchange of HRTFs in a binaural rendering stage was found to be significantly worse.

Rotation Methods	t-test
TOA - FOA	t(246) = 4.09; p<0.001
TOA - HRTF	t(232) = 8.08; p<0.001
FOA - HRTF	t(260) = 3.85; p<0.001
VBAP - HRTF	t(259) = 5.71; p<0.001
FOA - VBAP	t(263) = -1.93; p=0.054
TOA - VBAP	t(248) = 1.95; p=0.052

Table 1: Results of the t-tests.

It should be clear that more advanced methods of rotation and binaural rendering could affect the results. However, this study focused on state-of-the-art technology and available plugins. Therefore, it can be used as a guideline for implementation of rotation in virtual auditory environments. In fact, implementations of Ambisonics rendering, such as Google Omnitone make use of sound field rotation by matrix multiplication in Ambisonics domain.
The study was limited to binaural rendering in free field. Adding reverberation or rendering complete room impulse responses could influence the results and should be studied further. However, diffuse sound is not dependent on direction and would thus not be affected by the rotation. It may rather mask the artefacts due to rotation to some extent. In addition, a higher resolution HRTF database and more accurate rotation methods could improve the perceptual quality of the rotation in the binaural domain.
As a secondary result, it was found that listeners could perceive a significant difference between first order and third order Ambisonics representation. This supports the demand to use higher order Ambisonics formats in sound capturing, production tools and delivery chains.
It should be noted that the investigation was focused on Ambisonics input signals. There are known limitations of lower-order Ambisonics, such as limited spatial resolution. As stated, there are other common audio formats such as channel-based mixes and object-based audio that could benefit from other approaches for rotation. Their perceptual effects and complexity should be studied further to yield a more complete picture of optimal rotation methods.

References

[1] 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, Berlin*, May 2017.

[2] W. Heß, "Vergleich von Head-Tracking Systemen für virtuelle Akustik Applikationen," in *Fortschritte der Akustik - DAGA 2011*, pp. 677–678, Berlin: DEGA e.V., 2011.

[3] F. Zotter, "Analysis and Synthesis of Sound-Radiation with Spherical Arrays," PhD Thesis, University of Music and Performing Arts Graz, Austria, 2009.

[4] H. Møller, "Fundamentals of Binaural Technology," *Applied Acoustics*, vol. 36, pp. 171–218, 1992.

[5] J. Ivanic and K. Ruedenberg, "Rotation Matrices for Real Spherical Harmonics. Direct Determination by Recursion," *The Journal of Physical Chemistry*, vol. 100, no. 15, pp. 6342–6347, 1996.

[6] V. Pulkki, "Generic Panning Tools for MAX/MSP," in *Proceedings of International Computer Music Conference*, pp. 304–307, 2000.

[7] IRCAM Spat. `http://forumnet.ircam.fr/product/spat-en/`, 2017.

[8] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102, 2001.

[9] ambiX v0.2.7 - Ambisonic plug-in suite. `http://www.matthiaskronlachner.com/?p=2015`, 2017.

[10] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "Ambix – Suggesting an Ambisonics Format," in *3rd International Symposium on Ambisonics and Spherical Acoustics*, 2011.

[11] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.

[12] S. S. Sawilowsky, "New Effect Size Rules of Thumb," *Journal of Modern Applied Statistical Methods*, vol. 8, pp. 597–599, 2009.