applied sciences

MDPI

*Article*

# Auditory Localization in Low-Bitrate Compressed Ambisonic Scenes

**Tomasz Rudzki [1,*], Ignacio Gomez-Lanzaco [1], Jessica Stubbs [1], Jan Skoglund [2], Damian T. Murphy [1] and Gavin Kearney [1]**

[1] AudioLab, Communication Technologies Research Group, Department of Electronic Engineering, University of York, York YO10 5DD, UK

[2] Google, San Francisco, CA 94105, USA

* Correspondence: tr837@york.ac.uk

Received: 15 May 2019; Accepted: 26 June 2019; Published: 28 June 2019

**Abstract:** The increasing popularity of Ambisonics as a spatial audio format for streaming services poses new challenges to existing audio coding techniques. Immersive audio delivered to mobile devices requires an efficient bitrate compression that does not affect the spatial quality of the content. Good localizability of virtual sound sources is one of the key elements that must be preserved. This study was conducted to investigate the localization precision of virtual sound source presentations within Ambisonic scenes encoded with Opus low-bitrate compression at different bitrates and Ambisonic orders (1st, 3rd, and 5th). The test stimuli were reproduced over a 50-channel spherical loudspeaker configuration and binaurally using individually measured and generic Head-Related Transfer Functions (HRTFs). Participants were asked to adjust the position of a virtual acoustic pointer to match the position of virtual sound source within the bitrate-compressed Ambisonic scene. Results show that auditory localization in low-bitrate compressed Ambisonic scenes is not significantly affected by codec parameters. The key factors influencing localization are the rendering method and Ambisonic order truncation. This suggests that efficient perceptual coding might be successfully used for mobile spatial audio delivery.

**Keywords:** spatial audio; perceptual evaluation; listening tests; ambisonics; binaural; bitrate compression; auditory localization; audio codec; opus; streaming

## 1. Introduction

Immersive audio technology is an inevitable element of modern digital media. It is present in cinematic, music and installation arts, broadcast, computer games, virtual reality, and augmented reality applications. With the rise of 5G mobile networks, it is also expected to become a key element of communication services. Typical use case scenarios of using immersive audio in mobile technologies require binaural playback to spatialize the sound. For example, in *mobile VR*, where a mobile device is attached to the VR headset (e.g. Samsung Gear VR, https://www.samsung.com/global/galaxy/gear-vr), or both are integrated (e.g. Oculus Quest, https://www.oculus.com/quest), spatial audio is delivered through headphones or miniature speakers built into the headset. Some recently introduced headphone products enable the use of motion sensors paving the way for interactive audio rendering without the visual presentation (e.g. Bose Frames, QC35, https://www.bose.com/en_us/better_with_bose/augmented_reality.html). With the current state of technology these wearable products require an external mobile device working in tandem, acting as a real-time audio processing unit. Such products might soon enhance general use navigation apps or help improve the accessibility for blind and visually impaired users. The recent introduction of immersive audio streaming for films and television programs for home cinema setups also indicates the need for mobile-based spatial audio solutions.

In addition, mobile spatial audio might become a large-scale delivery medium for music, superseding the existing formats originally intended to use multiple loudspeakers for audio playback.

State-of-the-art immersive audio rendering systems are expected to give users the sensation of being in another acoustical space, as well as realistically render virtual sound sources as they would exist in the real world. To create a convincing auditory experience a set of sonic attributes needs to be provided, e.g. natural (or plausible) timbre, sound externalization, convincing acoustics, and precise sound source localization. For high quality immersive and interactive experiences, the accurate presentation of virtual sound source location is critical as it is required for directing user's attention and providing coherent visual and auditory cues.

Key external factors affecting the auditory localization in immersive audio reproduction include: Head-Related Transfer Function (HRTF) mismatch, the frequency response of headphones or loudspeakers, and audio rendering method. For the latter consideration, one of the main techniques which is employed in mobile spatial audio rendering is Ambisonics [1]. It is used by content delivery services (e.g. YouTube, Facebook) to stream spatial audio for 360/180-degree videos. Current perceptual audio coding standards officially supporting Ambisonics are MPEG-H 3D Audio [2] and Opus [3]. However, such perceptual audio coding can also introduce spatial distortions leading to degradation of localization cues. This is the primary focus of this paper—the evaluation of auditory localization within Ambisonic scenes with respect to perceptual low-bitrate coding and different reproduction methods. Specifically, the subjective differences between Ambisonic scenes encoded with Opus at different bitrates and Ambisonic orders are investigated in terms of localization precision of virtual sound sources in loudspeaker and headphone-based presentations.

This paper is organized as follows. Firstly, a review of the research on human auditory localization, binaural reproduction of Ambisonics, and localization performance in spatial audio systems is presented in Section 2. Section 3 then details the methodology for the described experiment, along with results presented in Section 4. The results are discussed and planned future work is described in Sections 5 and 6.

## 2. Background

### 2.1. Human Auditory Localization

The basic property of spatial hearing is the ability to derive information about the sound source location based on the phase and level differences in left and right ear signals. Additional sound source localization information is provided by pinna-related spectral cues. An in-depth review of published research on human auditory localization can be found in [4,5]. A concise summary is provided here. The performance of human auditory localization varies depending on the direction of incidence of the sound wave, distance, level, and frequency characteristics of the sound source. The misplacement between the perceived location of the sound source and its actual location introduces a *constant localization error*, which can be interpreted as the *accuracy* of auditory localization. The variability of listener's perception and listening conditions contributes to the *random localization error*, which represents the *precision* of auditory localization. This concept of localization accuracy and precision is illustrated in Figure 1.

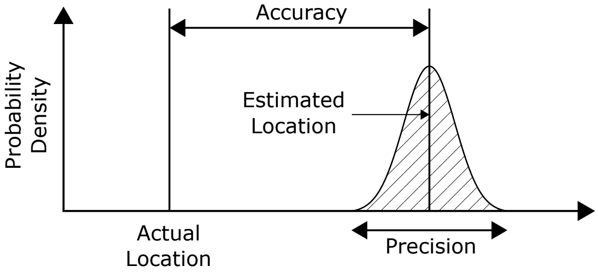

**Figure 1.** Auditory localization accuracy and precision. Adapted from [5].

The directional spatial resolution (precision) of the auditory system can be associated with the minimum difference in the direction of the sound source that causes a change in the perceived position of an auditory event. Blauert describes this attribute as *directional auditory localization blur*. It is also referred to as Minimum Audible Angle (MAA), when obtained in sound-source-discrimination experiments. The lower limit of the MAA in the horizontal plane is about 1° [6,7] and it is observed for sound sources located in front of the listener. The horizontal MAA increases to about 10° for the lateral directions. This confirms that spatial auditory resolution in the horizontal plane depends on the discrimination thresholds of interaural time and level differences, which change more rapidly in the azimuth for sources located ahead of the listener in comparison to the lateral region.

The reported MAA threshold in the median plane for sound sources placed in front of the listener is approximately 3° [7]. In this case, the auditory localization performance depends on the directional filtering of the pinna and body. For sources located on the diagonal planes spatial resolution depends on both interaural differences and spectral changes [7,8]. Auditory localization performance degrades in the presence of other sound sources [9].

### *2.2. Binaural-Based Ambisonics*

If the binaural localization cues (ITD, ILD) and spectral cues are reproduced correctly, the perceived spatial and timbral attributes of presented sound should be the same in comparison to the real sound source. This can be achieved through the controlled headphone-based reproduction of binaural signals. These signals can be recorded using binaural microphones or artificially synthesized employing HRTF sets. Readers are referred to [10–12] for more complete references on binaural recording technique and synthesis.

HRTF-based signal processing is a key element of all interactive binaural reproduction systems, as it can accommodate head-tracking, enabling dynamic spatial audio rendering using information on the listener's head movements. If the signals fed to each ear are modified accordingly, providing localization cues based on the head displacement and rotation, presented virtual sound sources will remain at their positions in space. According to [13,14] perception of virtual sound sources is dramatically improved with head-tracking enabled systems. The simplest head-tracking devices provide information on the rotation of the listener's head. An accurate approximation of the head orientation can be derived using inertial and magnetic sensors. To track head displacement, usually additional optical tracking is required.

An alternative to direct HRTF convolution is to use HRTFs as virtual loudspeakers. Here the same signals that would be used in a real-world 3D loudspeaker array are convolved with the HRTFs corresponding to specific loudspeaker positions to give a virtualized presentation of the array over headphones. The accuracy of sound field reproduction then becomes dependent on the spatialization method used, for example, Vector-Base Amplitude Panning [15], Wave Field Synthesis [16] or Ambisonics [1]. The latter spatialization method is an approach to deliver an approximation of the full-sphere sound field at the listener's ears. The theory of Ambisonics is well documented and the reader is directed to [17–19] for good explanations of the topic. A succinct review is provided here.

Ambisonics is based on the spherical harmonic representation of the sound field. A sound source can be encoded to Ambisonics format through matrix multiplication with spherical harmonic weights that represent the source position on the sphere. Ambisonic sound field representations employing 0th and 1st-order spherical harmonics are known as First-Order Ambisonics (FOA). Sound field representations that extend beyond first order are referred to as Higher-Order Ambisonics (HOA).

In practice, a limited number of Ambisonic components can be transmitted and exploited. The required number of components $N$ for a periphonic system of $m$th order can be calculated by $N = (m+1)^2$. Increasing the order results in a higher spatial aliasing frequency $f_{\text{alias}}$ and results in a larger sweet spot [18]. A limited number of spherical harmonic components leads to truncated representation of the sound field and a decrease in spatial resolution. Therefore, to achieve higher spatial resolution, higher orders must be used. Above $f_{\text{alias}}$ Ambisonic decoding introduces timbral alterations to the original encoded signals [20] degrading overall timbral fidelity.

Ambisonics has been widely adopted for immersive applications for mobile media since it can be easily manipulated and transformed [21] to facilitate stable sound sources when using dynamic binaural rendering. In this case, the Ambisonic scene is rotated counter to head movements by scaling the Ambisonic channels by correct coefficients adjusted according to the head-tracking data. Please note that the directions of virtual loudspeakers (simulated by HRTFs) remain unchanged and it is the loudspeaker feeds to the virtual loudspeakers that are updated in real time. Figure 2 shows the basic audio signal chain for loudspeaker-based and binaural-based Ambisonic rendering.

Ambisonics also allows for a reduction of bandwidth and computational requirements needed for delivering immersive audio content in comparison to the traditional multichannel surround formats and object-based approaches [22]. Most of the binaural rendering systems supporting Ambisonics use computationally efficient spherically decomposed HRTF sets, i.e., pre-computed combinations of virtual loudspeakers and HRTFs [23]. Several HRTF manipulation techniques have been proposed for optimal binaural rendering of Ambisonic signals [20,24,25].

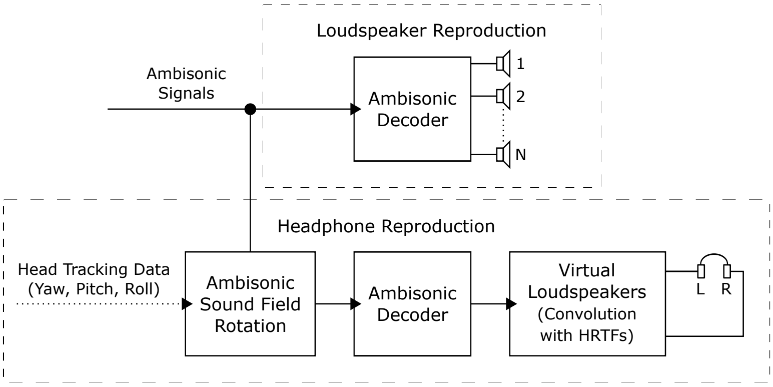
Loudspeaker Reproduction
Ambisonic Signals
Ambisonic Decoder
1
2
N
Headphone Reproduction
Head Tracking Data (Yaw, Pitch, Roll)
Ambisonic Sound Field Rotation
Ambisonic Decoder
Virtual Loudspeakers (Convolution with HRTFs)
L
R

**Figure 2.** Block diagram illustrating audio signal chain for Ambisonic rendering over loudspeakers and headphones.

### 2.3. *Auditory Localization in Ambisonics*

Auditory localization in spatial audio systems can be evaluated by measuring human performance in sound source localization tasks. Such indirect tests use various response techniques, e.g. perceived direction reporting, visual mapping, physical pointing, and acoustic pointer adjustment. Egocentric pointing methods employ localization judgment reported in a coordinate system centered on the listener's body. The position of the sound source is indicated by the listener's hand or head pointed towards the sound source [26]. Another type of egocentric method is *proximal pointing*, where the listener points in the apparent direction of the sound source in the proximal region of the head [27]. In exocentric methods, the listener indicates the perceived direction of the sound source using an external device, e.g. pointing with a stylus on a solid sphere placed in front of their body [28]. Another method of localization performance evaluation employs the use of a real or virtual acoustic pointer. The practical implementation of this method has been described in Section 3.

The relation between Ambisonic order and auditory localization error has been researched in several experiments. Previous studies used synthesized Ambisonic scenes as well as recorded with Ambisonic microphones [29,30] reproduced using loudspeaker arrays [31,32] and binaural rendering [33]. It has been shown that localization error depends on the Ambisonic order as well as the incidence of the virtual sound source. Furthermore, binaural reproduction produces more front-back confusion errors in comparison to loudspeaker-based reproduction [33]. Both loudspeaker and headphone reproduction methods have not been directly compared using HOA and individual HRTF-based rendering.

There is currently a limited amount of research published on the quality of compressed spatial audio, and particularly, on the compression of First and HOA. The recent version of the Opus codec (https://people.xiph.org/~jm/opus/opus-1.3) implements Channel Mapping Family 3 which allows for Ambisonic signal coupling [34]. Previous work by Narbutt et al. includes subjective evaluation of Ambisonics compressed with Opus 1.2 codec with Channel Mapping Family 2 implementation [35] and the development of a reference objective spatial audio quality metric [36]. They use a MUSHRA paradigm to assess the localization degradation and demonstrate quality degradation between equivalent bitrates at different orders. The absolute extent of localization precision is not shown. These studies also focus on static and generic HRTF binaural listening conditions. The localization performance within Ambisonic scenes compressed with Opus Channel Mapping Family 3 has been researched using loudspeaker-based reproduction exclusively [32]. The work presented in this paper extends this research by binaural evaluation.

## 3. Methods

The purpose of the experiment presented in this paper was to subjectively assess the spatial distortion introduced by Ambisonic order truncation and perceptual coding of Ambisonic scenes using different bitrates. The method of adjustment [37,38] was used for the auditory localization tests. Participants were asked to move an artificially reproduced virtual acoustic pointer to the perceived direction of a reproduced target sound source using a physical controller [39] (see Figure 3). The audio playback of pointer and target scenes was controlled by participants and programmed to ensure that both stimuli were never presented simultaneously. The azimuth and elevation step encoders adjusted the rotation of the rendered acoustic pointer with a single-degree precision. It is important to note that the experiment was designed to examine the perceived differences between the uncompressed and low-bitrate compressed scenes, not the absolute localization error. Both target sound source and acoustic pointer were reproduced using the same rendering method. Assessing the relative localization of virtual sound sources removes the need for real sound sources used as the localization anchor. Therefore, evaluation of binaurally rendered signals reproduced using ear occluding headphones is possible.

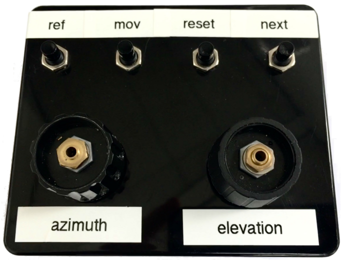

**Figure 3.** Physical controller designed for the auditory localization test.

### 3.1. Test Stimuli

The acoustic pointer consisted of a one-second pink noise burst encoded into 5th-order Ambisonics, pertinent to the spatial resolution of the Ambisonic reproduction systems used in the experiment. The low-bitrate compressed scenes presented during the simple scene evaluation consisted of one-second pink noise bursts placed in six static target directions: above, behind, and on the sides of the listener. The coordinates of investigated directions are listed in Table 1. These directions were chosen to match the context of the experiment—360-degree video streaming, where spatial audio is often used to direct user's attention in the virtual space.

The complex-scene stimuli consisted of the reference pink noise bursts and a modified Ambisonic soundscape [40] recorded with a 4th-order Ambisonic microphone (https://mhacoustics.com/products). The HOA microphone signal was used to provide different input signal conditions for the Opus codec in comparison to the simple scene material and to mimic the typical audio content of 360 videos. The used excerpt of the forest soundscape did not include any prominent spatially defined sounds which could influence the perception of the target sound direction. The level of the soundscape was empirically adjusted to prevent the participants from being significantly distracted from the task. Because the Ambisonic soundscape spatial resolution was limited to the 4th order, additional 5th-order background noise was added. This 5th-order noise signal was synthesized using 36 uniformly distributed virtual loudspeakers fed with decorrelated Brownian noise samples. Brownian noise was chosen because its power spectrum differs from the pink noise used as the virtual sound sources in this experiment. The target sound sources consisting of pink noise bursts (4 times repeated sequence of 2 s burst with 250 ms rise and fall times followed by 500 ms of silence) were panned at the specified directions shown in Table 1.

**Table 1.** Reference sound source directions during the localization performance test.

| **Direction** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| Azimuth (°) | 0 | 180 | 72 | 324 | 216 | 108 |
| Elevation (°) | 90 | $-18$ | 18 | $-18$ | 18 | $-18$ |

The required 1st and 3rd-order test stimuli were extracted from the 5th-order simple and complex scenes and subsequently compressed using Opus encoder at different bitrates and with Channel Mapping Family 3 enabled. The resulting test stimuli set consisted of 60 scenes for both simple and complex-scene tests (a multiplication of six target sound source directions and ten system conditions). The investigated system conditions are shown in Table 2. The uncompressed 5th-order Ambisonic scenes were used as the reference condition.

**Table 2.** Investigated bitrates (kbps) at different Ambisonic orders.

| | **Bitrate per Channel** | **Total Bitrate** | | |
|---|---|---|---|---|
| | | **1OA** | **3OA** | **5OA** |
| Compressed | 16 | 64 | 256 | 576 |
| Compressed | 32 | 128 | 512 | 1152 |
| Compressed | 64 | 256 | 1024 | 2304 |
| Uncompressed | 768 | | | 27,648 |

## *3.2. Spatial Audio Rendering*

Evaluation was conducted using multi-loudspeaker and dynamic binaural rendering methods inside an acoustically treated room. The loudspeaker reproduction was done using a 50-channel full-sphere array based on the Lebedev quadrature [41], see Figure 4. The sound pressure level (SPL) at the center of the array was aligned for each individual loudspeaker using an automated SPL calibration script. The reproduction SPL in the center of the array was set to 65 dBA. The complex-scene stimuli levels were aligned subjectively to match the simple scene test loudness. The rendering of Ambisonic scenes was done using three different loudspeaker configurations: octahedron, 26-point Lebedev grid, and 50-point Lebedev grid, as optimal configurations for the 1st, 3rd and 5th-order Ambisonics signals respectively [41]. Dual-band decoding was implemented by pre-filtering the Ambisonic input with a set of shelf filters (https://github.com/resonance-audio/resonance-audio/tree/master/matlab/ambisonics/shelf_filters) and applying Max-Re correction weightings to the high-passed signals before feeding the decoder. The AmbiX (https://matthiaskronlachner.com/?p=2015) Ambisonic decoder configuration files were obtained from the SADIE II database (https://york.ac.uk/sadie-project/ambidec.html). Listening test software for loudspeaker presentation was created using the visual audio programming environment Max (https://cycling74.com/products/max).

To create binaural signals, loudspeaker feeds were convolved in real time with diffuse-field equalized HRTF sets obtained from the SADIE II database. Individual and generic HRTF sets were used. The individual HRTF-based evaluation required a participation of subjects who took part in the database creation. The generic HRTF evaluation was done using an HRTF set obtained using Neumann KU100 binaural microphone. Sennheiser HD 650 headphones were used for the binaural tests due to the low variability in frequency response between coupling and decoupling of the headphones with the ears [42]. Frequency response of the headphones was compensated using inverse filters based on responses measured with the KU100 dummy-head. Binaural reproduction level was adjusted to match the loudspeaker reproduction level through calibration with a KU100 head. An Optitrack optical motion tracking system (https://optitrack.com) using six Flex-3 infrared cameras and reflective markers attached to the headphone headband was used for dynamic binaural rendering. Headphone-based tests were conducted using dedicated listening test software [39] and the DAW Reaper (https://reaper.fm) as the audio engine.

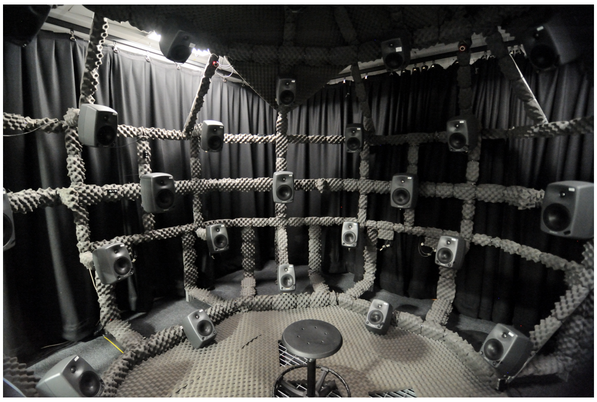

**Figure 4.** 50-channel spherical loudspeaker array at the AudioLab, University of York.

### *3.3. Participants*

The experimental group consisted of MSc and PhD audio engineering students as well as senior researchers with experience in critical listening. Some of the participants took part in the sound quality assessment tests for the first time. All participants were instructed how to perform the tests by reading an information sheet and receiving individual demonstrations. The localization test included a training phase consisting of three test tasks with uncompressed stimuli. The responses gathered during the training were not exported for further analysis. Participants were instructed to keep their heads at the center of the loudspeaker rig and limit head movements throughout the test, although their heads were not physically constrained. All subjects gave their informed consent for inclusion before they participated in the study. The protocol was approved by the Physical Sciences Ethics Committee of the University of York (approval code: Rudzki021018).

## **4. Results**

The collected directional data represents auditory localization of the virtual acoustic pointer adjusted to match the perceived direction of the virtual target sound sources. The responses were gathered during 104 listening tests consisting of 60 individual tasks each. Table 3 shows the number of participants who completed the tests, separated into subgroups by the reproduction method and audio content type used. The median time of completing each individual task by the participants was about 26 s. All the investigated low-bitrate compression conditions were compared within each of the six subgroups.

**Table 3.** Number of participants who completed the tests grouped by the rendering method and audio content type used.

| **Reproduction Method** | **Loudspeakers** | | **Binaural (Individual HRTFs)** | | **Binaural (Generic HRTFs)** | |
|---|---|---|---|---|---|---|
| Content Type | Simple | Complex | Simple | Complex | Simple | Complex |
| Number of Participants | 21 | 16 | 15 | 14 | 19 | 19 |

Figure 5 shows the distribution of the acoustic pointer directions on the sphere set by participants. Each sphere represents data gathered for different directions of the target sound source presented using all three reproduction methods at different Ambisonic orders and compression bitrates. Figure 6 shows the same data plotted using equirectangular projection.

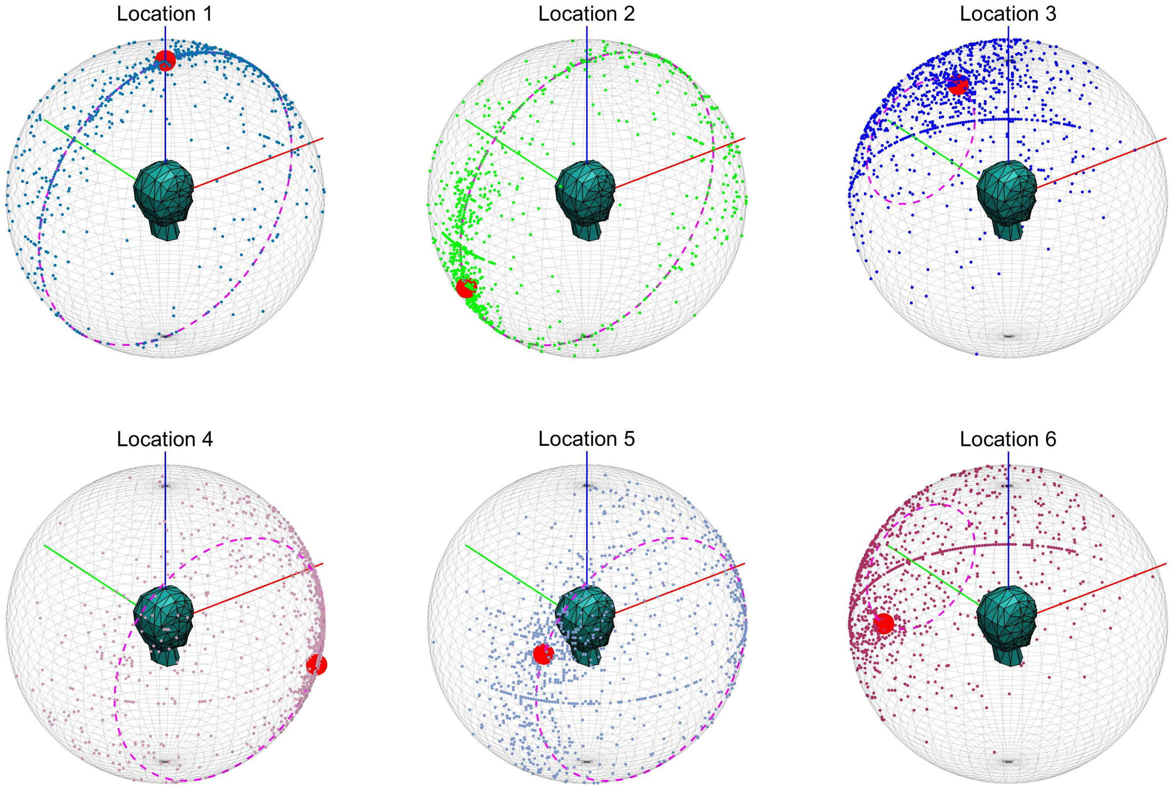

**Figure 5.** Distribution of the acoustic pointer direction recorded during the listening test corresponding to each target source direction. The red dots symbolize directions of the target sound sources. The dashed magenta circles represent the respective cones of confusion on the sphere. The axes denote directions relative to the listener: red—front, green—left, blue—top.

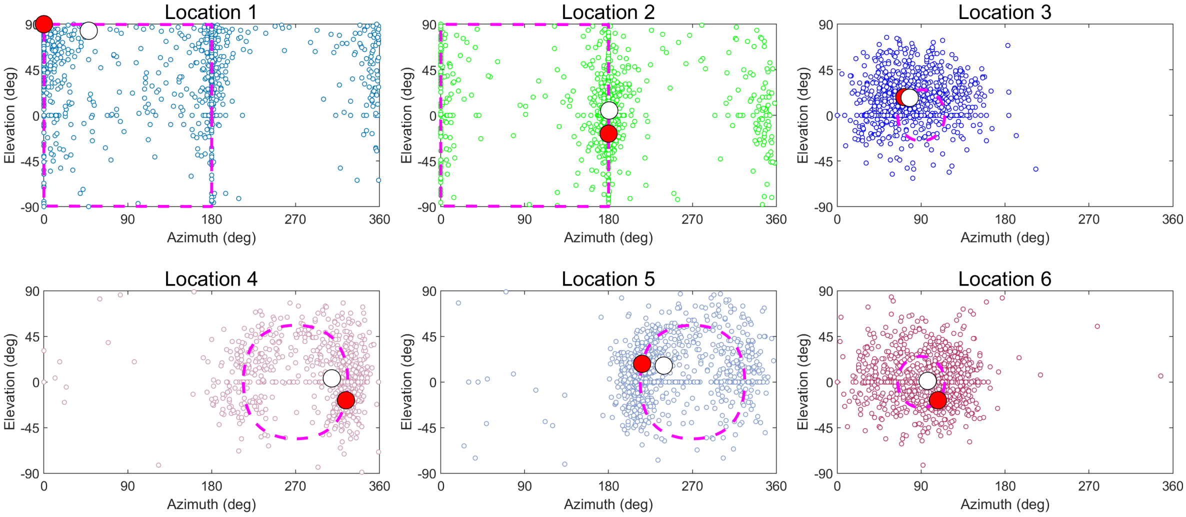

**Figure 6.** Distribution of the acoustic pointer directions recorded during the listening test corresponding to each target source direction. The red dots symbolize directions of the target sound sources. The dashed magenta lines represent the respective cones of confusion. The unfilled circles represent mean direction of the pointer directions.

The initial pointer direction for each task was set straight in front of the listener. Since participants operated the azimuth and elevation controls (see Figure 3) independently, it can be seen that the distributions of recorded pointer directions are slightly skewed towards the horizontal and median planes. This suggests that responses may have been affected by the collection method. Directions one and two of the target virtual sound source correspond to the median plane directions of incidence. The respective recorded acoustic pointer indications are distributed close to the intersection of the median plane with the unity sphere. Perceived elevation of these sources and acoustic pointer can be matched by the listener using spectral and dynamic localization cues exclusively. Directions four and five correspond to slightly elevated and laterally shifted directions. Collected acoustic pointer indications are distributed along the intersection of respective cones of confusion with the unity sphere. Directions three and six correspond to slightly elevated and strongly laterally shifted directions, close to the interaural axis. Based on the visual observation, the distributions of pointer indications are rather concentrated around the target sound source directions with a slight skew towards the respective cones of confusion.

Further analysis was conducted using great-circle distance, which can be calculated as the shortest angular distance between each pointer and corresponding target directions on the unity sphere. The analysis of horizontal and vertical localization error components was performed; however, the exhibited differences between codec conditions were less significant than when using the combined error metric. To minimize the directional bias introduced by the pointing interface and focus on the random localization error, the great-circle distance was calculated using the spherical means as the reference directions, not the encoded target directions. Mean spherical direction was calculated for each of the analyzed subsets.

Figure 7 shows the set of probability density functions [43] of the localization error obtained experimentally at different Ambisonic orders/codec bitrates and rendering methods. It can be seen that the general shape of presented distributions corresponds to the shape of von Mises-Fisher distribution [44] plotted as the probability density function of the distance between each spherical mean and each sample. However, the experimental distributions exhibit multimodal characteristics caused by the cone of confusion and data collection biases. This limits the use of statistical tests based on parameterized spherical data distributions for a unified analysis of the results. Instead, the Kruskal-Wallis rank-based non-parametric test was used to investigate the spherical concentration [45,46] of the participant responses under different experimental conditions.

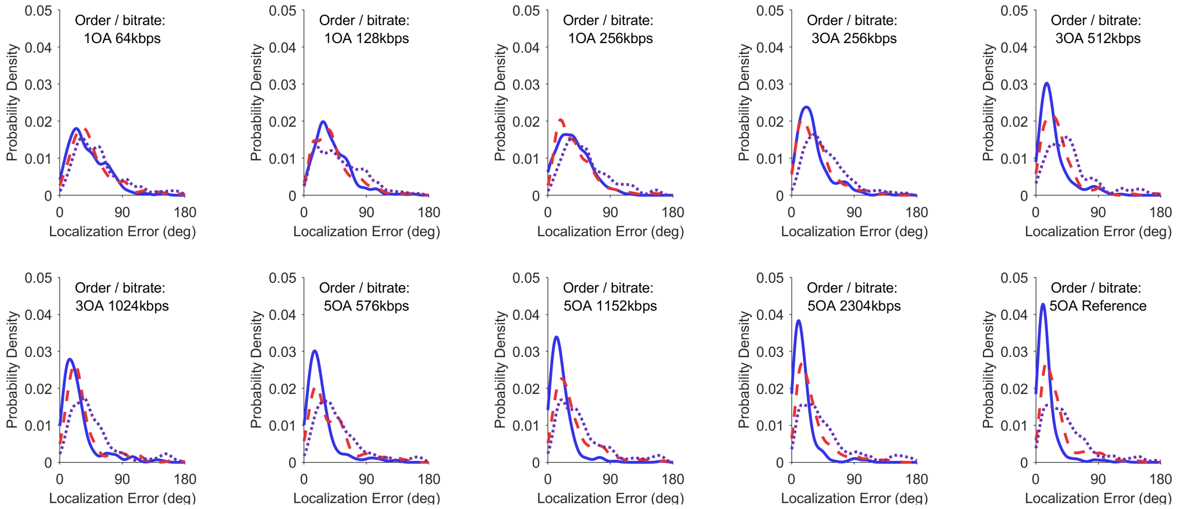

**Figure 7.** Probability density functions of the localization error for different spatial audio rendering methods at different codec bitrates and Ambisonic orders. Continuous line represents loudspeaker reproduction, dashed line—binaural (individual HRTFs), dotted—binaural (generic HRTFs). Kernel bandwidth: $BW = 6$. HRTFs: Head-Related Transfer Functions.

The following experimental variables were tested: participants, virtual target sound source direction, codec bitrates, Ambisonic orders, audio content type, and audio reproduction method (loudspeaker-based vs. individualized vs. generic HRTFs). A significant difference in the overall localization task performance between participants has been found in each of the three experimental phases using different spatial audio rendering methods ($p < 0.01$). The overall localization error median for each participant is shown in Figure A1.

The effect of the position of the virtual target sound source was significant ($p < 0.01$); however, no clear trends in data were identified. The localization error median for each direction at different rendering methods is shown in Figure A2. The effect of codec bitrate was analyzed in nine subgroups, grouped by Ambisonic order and rendering method. It was found to be significant ($p < 0.01$) in two groups: 3rd-order and 5th-order scenes reproduced using the loudspeaker array. Significant differences between compressed scenes grouped by Ambisonic order have been found for each of the three rendering methods ($p < 0.01$). The effect of content type on participant responses was investigated in the raw data and the test result was close to the 95% confidence limit ($\chi^2 = 3.92, p = 0.048$). Detailed analysis was done in 30 subgroups, grouped by different codec bitrates / Ambisonic orders and spatial audio rendering methods. The difference between simple and complex-scene content has not been found to be significant ($p > 0.01$) in 29 of the 30 subgroups. Rendering method had significant effect on the localization error ($p < 0.01$).

Figure 8 shows median localization error at different codec bitrates, Ambisonic orders, and rendering methods. It can be seen that generic HRTF reproduction resulted in higher localization error compared to the loudspeaker and individual HRTF reproduction. A decrease in localization error with the increase of Ambisonic order was observed using all three reproduction methods. This effect is most prominent at the loudspeaker-based tests, where the difference between 1st and 3rd-order is much more significant than the difference between 3rd and 5th-order. The differences in median localization error caused by different bitrates within the same Ambisonic order can be observed; however, they are not significant in most cases. Based on the obtained results we can infer that the localization precision depends slightly on the bitrate (within the examined bitrate values).

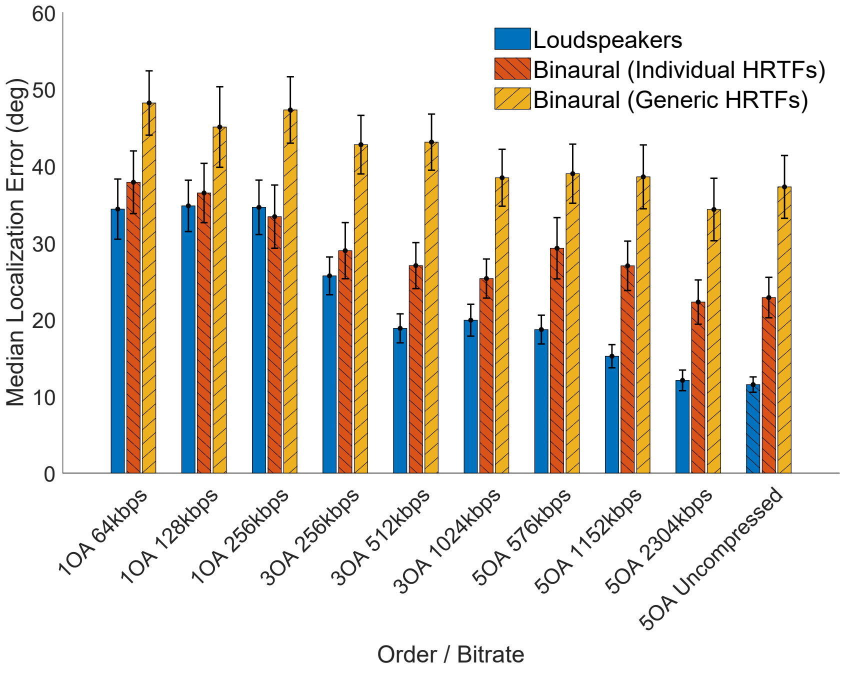

**Figure 8.** Median localization error at different codec bitrates/Ambisonic orders. The whiskers indicate non-parametric 95% confidence intervals [47].

## 5. Discussion

Multiple factors have contributed to the localization error measured during the experiment. Firstly, the limited resolution of the human auditory localization in 3D space, which has been briefly described in Section 2. This study was focused on directions where the localization blur is relatively high, which may have contributed to the high variance in the responses. The measured median localization error for the 5th-order uncompressed reference reproduced over loudspeaker array was about 11°, which is comparable with other studies focusing on localization in horizontal and vertical planes combined [48].

Secondarily, the limited-order Ambisonic representation of the target and acoustic pointer scenes has contributed to the localization error. The results prove that localization precision is largely defined by the Ambisonic order, as higher orders present more precise spatial resolution. The median localization error for the 1st-order scenes was about 34°what corresponds to the results obtained by Braun [29] for the virtual sound source presented over the loudspeaker array. The localization error obtained in the same study for the 4th-order Ambisonic virtual sources was about 10°, what corresponds indirectly to our 5th-order uncompressed condition result. In the experiment by Bertet [30], the localization error for 4th-order Ambisonic virtual sources measured in the horizontal plane varied from 5° to 14° at the lateral positions. It is important to note that our experiment used virtual sound source presentation for both target and pointer sounds to facilitate the headphone-based tests, where the study by Bertet was done with an Ambisonic pointer and real sound source target. A direct comparison of our results with any of the referenced studies is not possible due to the differences in reproduction systems and test frameworks.

Another source of error comes from the limitations of the reproduction methods used. Participants who took part in loudspeaker tests localized the sound sources with the highest precision. A similar degree of precision was obtained using binaural reproduction with individually measured HRTFs at 1st Ambisonic order. At 3rd and 5th-order, the localization error in headphone-based tests was higher than in the loudspeaker-based phase. Binaural reproduction of Ambisonic scenes employing the generic HRTF set resulted in the highest localization error at all tested signal conditions. This phenomenon requires further investigation, as the loudspeaker and headphone test data were obtained with different groups of participants. It is worth noting that the lowest bit rate in 3rd-order Ambisonic presentations (256 kbps) produced a significantly improved localization precision in the loudspeaker case than the highest bit rate condition for 1st order (also 256 kbps). In both headphone listening cases there is no significant difference between the aforementioned bitrates and orders. These results are contrary to those found by Narbutt et al. which show a significant degradation with the lowest bitrate at 3rd order [36]. We attribute these differences to the use of head-tracking and the greater localization-performance test paradigm of this study rather than perceived overall quality.

The results of this study show that auditory localization in low-bitrate compressed Ambisonic scenes is not significantly affected by codec parameters. Although the differences between localization error for the same orders are not statistically significant, based on visible trends it can be seen that localization precision does degrade slightly with lower bitrate compression. The other studies focused on timbral fidelity of the Opus compressed spatial audio revealed significant differences between bitrates and Ambisonic orders [32].

The effect of an additional soundscape present in test stimuli was investigated; however, no significant difference was observed between simple and complex content presentations. The 5th-order diffused sound scene was added to investigate if the spatial distortion of the single sound source presentation within the scene will be affected by feeding additional non-directional information to each of the encoded channels. This condition was supposed to mimic a recording done with an Ambisonic microphone, although maintaining the highest possible spatial resolution of the single sound source by synthesizing the Ambisonic sound field. The impact of the sound scene complexity on the localization error has not been revealed.

The chosen data collection method might have contributed to a high variance in the participants' overall localization performance. The average duration of the test session was about 45 minutes with a single break in between. Some of the participants reported a mild psycho-physical fatigue after the experiment which suggests that responses collected using a less challenging test methodology could give more consistent results across participants. As the participants' heads were not constrained and the head movements were not recorded by the optical tracking system, it is unknown to what degree the small head rotations also contributed to the measured localization performance.

Another factor affecting localization precision in the experiment is the acoustic pointer response collection method. It is possible that the auditory localization precision measured using the acoustic pointing method gives higher error estimates than the source discrimination methods used for the MAA measurements, which is focused more on the change of the perceived acoustic signal rather than spatial analysis of the sound field [48]. However, once the virtual pointer and target directions are perceptually matched, participants might compare both signals using other features than perceived spatial locations. The degree in which the presented mechanism has contributed to the experimental results remains unknown. Further studies should consider different response collection techniques adequate to the proposed application of the coding system.

The continuation of research will investigate the development of efficient indirect perceptual evaluation methods for the assessment of binaural-based spatial audio systems, including bitrate compression schemes. We will also investigate localization accuracy and precision for frontal presentations in more depth given the importance of this region for immersive content consumption in VR and AR as well as teleconferencing services.

## 6. Conclusions

Perceptual evaluation of bitrate compression schemes is an important part of mobile spatial audio technology development. Delivering spatially accurate and precise auditory information at low data bandwidth allows for a wider adaptation of immersive technologies. In this paper, we have presented a study on the localization precision of binaural-based Ambisonic reproduction using low-bitrate compression over different Ambisonic orders. The tests were conducted using headphone-based reproduction employing both individualized and generic HRTF sets and with real-world loudspeaker presentations for comparison. We conclude that using strong bitrate compression will not affect the auditory localization in scenes encoded using Opus compression when compared to uncompressed Ambisonic presentations; however, the timbral fidelity aspect of the compressed audio should be considered as well. Using HOA content instead of 1st-order Ambisonics will improve localization within the scenes, especially when using personalized binaural rendering or multi-loudspeaker reproduction.

**Author Contributions:** Conceptualization, T.R., J.S. (Jan Skoglund) and G.K.; Methodology, T.R., I.G.-L., and G.K.; Software, T.R. and I.G.-L.; Validation, T.R. and G.K.; Formal Analysis, T.R.; Investigation, T.R.; Resources, T.R.; Data Curation, T.R., I.G.-L., and J.S. (Jessica Stubbs); Writing—Original Draft Preparation, T.R.; Writing—Review & Editing, T.R., J.S. (Jessica Stubbs) and G.K.; Visualization, T.R.; Supervision, D.T.M. and G.K.; Project Administration, T.R., J.S. (Jessica Stubbs), and G.K.; Funding Acquisition, J.S. (Jan Skoglund) and G.K.

**Funding:** This research was jointly funded by Google as part of a Chrome Media Research Award and by the Engineering and Physical Sciences Research Council (EP/M001210/1).

**Acknowledgments:** The authors gratefully acknowledge the participation of the listening test subjects.

**Conflicts of Interest:** The authors declare no conflict of interest. The sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| SPL | Sound Pressure Level |
| ITD | Interaural Time Difference |
| ILD | Interaural Level Difference |
| HRTF | Head-Related Transfer Function |
| HRIR | Head-Related Impulse Response |
| MAA | Minimum Audible Angle |
| HOA | Higher-Order Ambisonics |
| VR | Virtual Reality |
| AR | Augmented Reality |
| DAW | Digital Audio Workstation |

## Appendix A

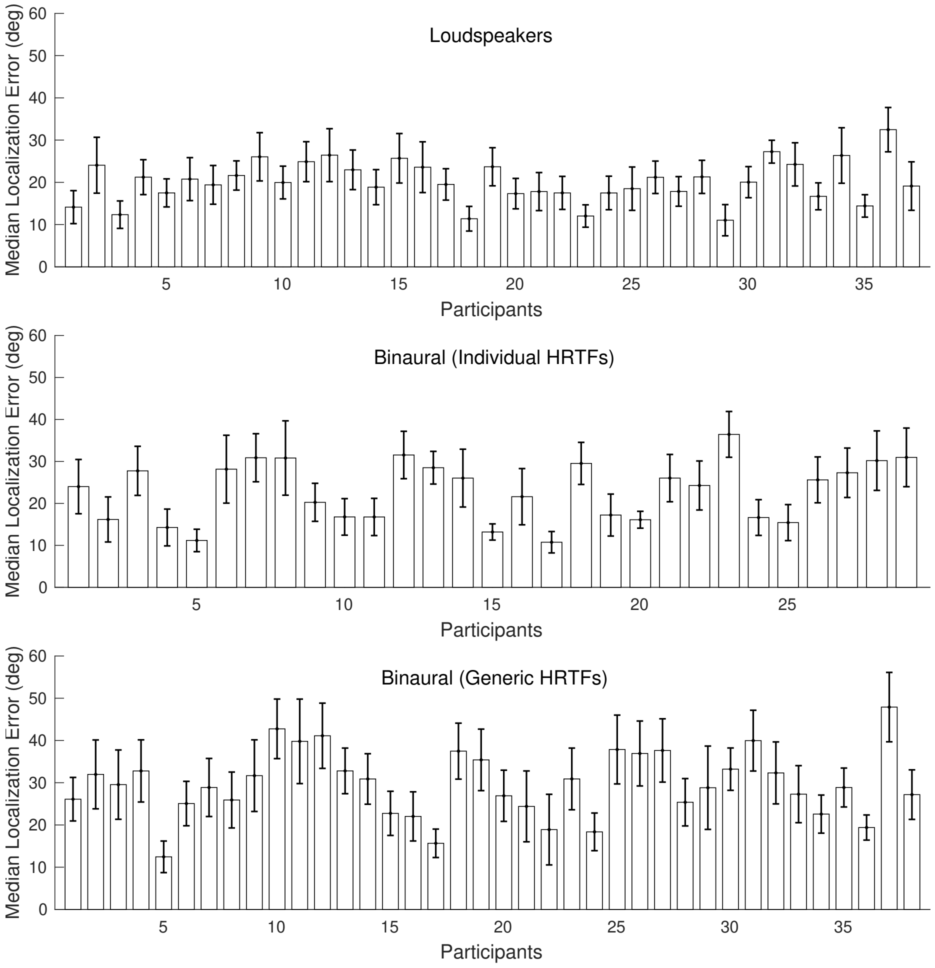

**Figure A1.** Median localization error of each participant at different reproduction methods. The whiskers indicate non-parametric 95% confidence intervals [47].

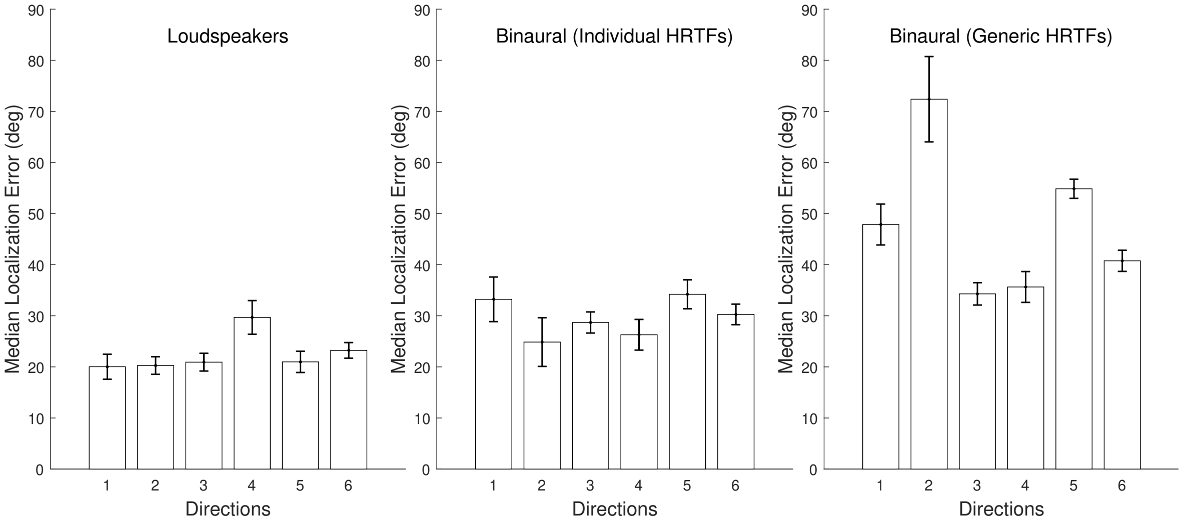

**Figure A2.** Median localization error for each virtual sound source direction at different reproduction methods. The whiskers indicate non-parametric 95% confidence intervals [47].

## References

1. Gerzon, M.A. Periphony: With-Height Sound Reproduction. *J. Audio Eng. Soc.* **1973**, *21*, 2–10.
2. Herre, J.; Hilpert, J.; Kuntz, A.; Plogsties, J. MPEG-H 3D Audio–The New Standard for Coding of Immersive Spatial Audio. *IEEE J. Sel. Top. Sign. Proces.* **2015**, *9*, 770–779. [CrossRef]
3. Valin, J.M.; Maxwell, G.; Terriberry, T.B.; Vos, K. High-Quality, Low-Delay Music Coding in the Opus Codec. *arXiv* **2016**, arXiv:1602.04845.
4. Blauert, J.; Allen, J. *Spatial Hearing: The Psychophysics of Human Sound Localization*; MIT Press: Cambridge, MA, USA, 1997.
5. Letowski, T.; T Letowski, S. *Auditory Spatial Perception: Auditory Localization*; No. ARL-TR-6016; U.S. Army Research Laboratory: Aberdeen Proving Ground, MD, USA, 2012.
6. Mills, A.W. On the Minimum Audible Angle. *J. Acoust. Soc. Am.* **1958**, *30*, 237–246. [CrossRef]
7. Perrott, D.R.; Saberi, K. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *J. Acoust. Soc. Am.* **1990**, *87*, 1728–1731. [CrossRef] [PubMed]
8. Grantham, D.W.; Hornsby, B.W.; Erpenbeck, E.A. Auditory spatial resolution in horizontal, vertical, and diagonal planes. *J. Acoust. Soc. Am.* **2003**, *114*, 1009–1022. [CrossRef] [PubMed]
9. Langendijk, E.H.; Kistler, D.J.; Wightman, F.L. Sound localization in the presence of one or two distracters. *J. Acoust. Soc. Am.* **2001**, *109*, 2123–2134. [CrossRef] [PubMed]
10. Wightman, F.L.; Kistler, D.J. Headphone simulation of free-field listening. I: stimulus synthesis. *J. Acoust. Soc. Am.* **1989**, *85*, 858–867. [CrossRef]
11. Møller, H. Fundamentals of binaural technology. *Appl. Acoust.* **1992**, *36*, 171–218. [CrossRef]
12. Begault, D.R.; Trejo, L.J. *3-D Sound for Virtual Reality and Multimedia*; NASA/TM-2000-209606; NASA Ames Research Center: Moffett Field, CA, USA, 2000.
13. Wightman, F.L.; Kistler, D.J. Resolution of front–back ambiguity in spatial hearing by listener and source movement. *J. Acoust. Soc. Am.* **1999**, *105*, 2841–2853. [CrossRef]
14. Begault, D.R.; Wenzel, E.M.; Anderson, M.R. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *J. Audio Eng. Soc.* **2001**, *49*, 904–916. [PubMed]
15. Pulkki, V. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc.* **1997**, *45*, 456–466.
16. Berkhout, A.J.; de Vries, D.; Vogel, P. Acoustic control by wave field synthesis. *J. Acoust. Soc. Am.* **1993**, *93*, 2764–2778. [CrossRef]

17. Zotter, F.; Frank, M. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*; Springer Topics in Signal Processing; Springer International Publishing: Cham, Switzerland, 2019.
18. Daniel, J.; Moreau, S.; Nicol, R. Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging. In *Audio Engineering Society Convention 114*; Audio Engineering Society: Amsterdam, The Netherlands, 2003.
19. Kearney, G. Auditory Scene Synthesis Using Virtual Acoustic Recording and Reproduction. PhD Thesis, Trinity College Dublin, Dublin, Ireland, 2010.
20. McKenzie, T.; Murphy, D.; Kearney, G. Diffuse-field equalisation of binaural ambisonic rendering. *Appl. Sci.* **2018**, *8*, 1956. [CrossRef]
21. Kronlachner, M. Spatial transformations for the alteration of ambisonic recordings. Master's Thesis, University of Music and Performing Arts, Graz, Institute of Electronic Music and Acoustics, Graz, Austria, 2014.
22. Brettle, J.; Skoglund, J. Open-Source Spatial Audio Compression for VR Content. In Proceedings of the SMPTE 2016 Annual Technical Conference and Exhibition, Los Angeles, CA, USA, 25–27 October 2016; pp. 1–9. [CrossRef]
23. Gorzel, M.; Allen, A.; Kelly, I.; Kammerl, J.; Gungormusler, A.; Yeh, H.; Boland, F. Efficient Encoding and Decoding of Binaural Sound with Resonance Audio. In Proceedings of the Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio, York, UK, 27–29 March 2019; Audio Engineering Society: New York, NY, USA, 2019; pp. 1–12.
24. McKenzie, T.; Murphy, D.T.; Kearney, G. Interaural Level Difference Optimization of Binaural Ambisonic Rendering. *Appl. Sci.* **2019**, *9*, 1226. [CrossRef]
25. Zaunschirm, M.; Schörkhuber, C.; Höldrich, R. Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint. *J. Acoust. Soc. Am.* **2018**, *143*, 3616–3627. [CrossRef]
26. Majdak, P.; Goupell, M.J.; Laback, B. 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Atten. Percept. Psychophys.* **2010**, *72*, 454–469. [CrossRef]
27. Bahu, H.; Carpentier, T.; Noisternig, M.; Warusfel, O. Comparison of different egocentric pointing methods for 3D sound localization experiments. *Acta Acust. Acust.* **2016**, *102*, 107–118. [CrossRef]
28. Gilkey, R.H.; Good, M.D.; Ericson, M.A.; Brinkman, J.; Stewart, J.M. A pointing technique for rapidly collecting localization responses in auditory research. *Behav. Res. Methods Instrum. Comput.* **1995**, *27*, 1–11. [CrossRef]
29. Braun, S.; Frank, M. Localization of 3D ambisonic recordings and ambisonic virtual sources. In Proceedings of the 1st International Conference on Spatial Audio, (Detmold), Detmold, Germany, 10–13 November 2011; pp. 1–6.
30. Bertet, S.; Daniel, J.; Parizet, E.; Warusfel, O. Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. *Acta Acust. Acust.* **2013**, *99*, 642–657. [CrossRef]
31. Power, P.; Davies, W.; Hirst, J.; Dunn, C. Localisation of elevated virtual sources in higher order ambisonic sound fields. In Proceedings of the Institute of Acoustics, Brighton, UK, 14–16 November 2012; pp. 1–14.
32. Rudzki, T.; Gomez-Lanzaco, I.; Hening, P.; Skoglund, J.; McKenzie, T.; Stubbs, J.; Murphy, D.; Kearney, G. Perceptual Evaluation of Bitrate Compressed Ambisonic Scenes in Loudspeaker Based Reproduction. In Proceedings of the Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio, York, UK, 27–29 March 2019; Audio Engineering Society: New York, NY, USA, 2019.
33. Thresh, L.; Armstrong, C.; Kearney, G. A Direct Comparison of Localization Performance When Using First, Third, and Fifth Ambisonics Order for Real Loudspeaker and Virtual Loudspeaker Rendering. In Proceedings of the Audio Engineering Society Convention 143, New York, NY, USA, 18–21 October 2017; Audio Engineering Society: New York, NY, USA, 2017.
34. Skoglund, J.; Graczyk, M. *Ambisonics in an Ogg Opus Container*; RFC 8486; Internet Engineering Task Force: Fremont, CA, USA, 2018. [CrossRef]
35. Narbutt, M.; O'Leary, S.; Allen, A.; Skoglund, J.; Hines, A. Streaming VR for immersion: Quality aspects of compressed spatial audio. In Proceedings of the 2017 IEEE 23rd International Conference on Virtual System & Multimedia (VSMM), Dublin, Ireland, 31 October–4 November 2017; pp. 1–6. [CrossRef]

36. Narbutt, M.; Allen, A.; Skoglund, J.; Chinen, M.; Hines, A. AMBIQUAL—A full reference objective quality metric for ambisonic spatial audio. In Proceedings of the IEEE 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), Cagliari, Italy, 29 May–1 June 2018; pp. 1–6.
37. Cardozo, B.L. Adjusting the Method of Adjustment: SD vs DL. *J. Acoust. Soc. Am.* **1965**, *37*, 786–792. [CrossRef]
38. Langendijk, E.H. Collecting localization response with a virtual acoustic pointer. *J. Acoust. Soc. Am.* **1997**, *101*. [CrossRef]
39. Rudzki, T.; Murphy, D.; Kearney, G. A DAW-Based Interactive Tool for Perceptual Spatial Audio Evaluation. In Proceedings of the Audio Engineering Society Convention 145, New York, NY, USA, 17–20 October 2018; Audio Engineering Society: New York, NY, USA, 2018; pp. 1–3.
40. Green, M.C.; Murphy, D. EigenScape: A database of spatial acoustic scene recordings. *Appl. Sci.* **2017**, *7*, 1204. [CrossRef]
41. Lecomte, P.; Gauthier, P.A.; Langrenne, C.; Berry, A.; Garcia, A. A Fifty-Node Lebedev Grid And Its Applications To Ambisonics. *J. Audio Eng. Soc.* **2016**, *64*, 868–881. [CrossRef]
42. Adams, S.; Boland, F. On the distortion of binaural localization cues using headphones. In Proceedings of the IET Irish Signals and Systems Conference, Cork, Ireland, 23–24 June 2010; Institution of Engineering and Technology: London, UK, 2010; pp. 82–87.
43. Shimazaki, H.; Shinomoto, S. Kernel bandwidth optimization in spike rate estimation. *J. Comput. Neurosci.* **2010**, *29*, 171–182. [CrossRef] [PubMed]
44. Fisher, N.I.; Lewis, T.; Embleton, B.J. *Statistical Analysis of Spherical Data*; Cambridge University Press: Cambridge, UK, 1993.
45. Kruskal, W.H.; Wallis, W.A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [CrossRef]
46. Verdebout, T. On some validity-robust tests for the homogeneity of concentrations on spheres. *J. Nonparametr. Stat.* **2015**, *27*, 372–383. [CrossRef]
47. McGill, R.; Tukey, J.W.; Larsen, W.A. Variations of box plots. *Am. Stat.* **1978**, *32*, 12–16.
48. Makous, J.C.; Middlebrooks, J.C. Two-dimensional sound localization by human listeners. *J. Acoust. Soc. Am.* **1990**, *87*, 2188–2200. [CrossRef]