Audio Engineering Society

Convention Express Paper 241

Presented at the 156th Convention

2024 June 15-17, Madrid, Spain

Comparison of ambisonic and object-based spatial sound recording techniques

Patryk Kosior[1] and Bartłomiej Mróz[1]

[1] *Gdańsk University of Technology, Gabriela Narutowicza 11/12, 80-233, Gdańsk, Poland*

Correspondence should be addressed to Patryk Kosior (s180417@student.pg.edu.pl)

ABSTRACT

This article presents a comparison of spatial sound recording techniques based on scene-based and object-based audio. The study aimed to make different mixes from a recording which consists of a higher-order ambisonic microphone and spot microphones. For spot microphones simple ambisonics encoding was used, which allows panning the individual channels on an ambisonic sphere as objects. Recordings were combined in various variants of spatial resolution, mainly varying the order of ambisonics used. In the next step, a MUSHRA-like test was conducted on a panel of experts in auditory experiments. The experiment was done on headphones using a binaural rendering with three degrees of freedom provided via a head tracker. The findings suggest that the optimal immersion approach involved using individual object stimuli rendered at a 5th-order ambisonic spatial resolution. Regarding the ability to localize sounds, the combination of 3rd-order ambisonic with 5th-order objects yielded the highest performance. Overall, the outcomes of this experiment provide insights into recording and mixing techniques within the field of spatial audio.

1 Introduction

1.1 Ambisonics and scene-based audio

Ambisonics is a system for encoding and rendering a three-dimensional sound field [2]. Ambisonics is described by so-called ambisonics orders. Each order corresponds to a certain number of channels representing microphones with increasingly complex directivity patterns. Ambisonic orders can be represented as a collection of a certain number of microphones with given directional characteristics, symbolized by so-called spherical harmonics [26].

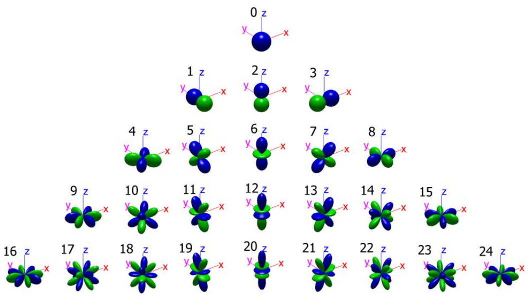


Figure 1. Spherical harmonics up to 4th order [26]

Higher ambisonic orders provide more complex directional patterns. This, in turn, makes it possible to increase the spatial resolution of the sound and a better rendering of the directivity of the sound coming to the microphone. Better directivity provides better localization of sound sources. However, the computational complexity also increases, which is important at the processing stage. The number of channels per order of ambisonics can be calculated from the formula: $(N+1)^2$, where N denotes the ambisonics order (in 3D ambisonics).

Ambisonics reflects the full sphere around the listener and is more commonly used in virtual reality applications – mainly because it is a coincidental technique. For high-fidelity recordings, multichannel-spaced microphone techniques (e.g., Fukada Tree, 2L Cube, or PCMA-3D) are more commonly used [6]. However, ambisonics can be decoded into various audio reproduction configurations (e.g., surround, stereo, or binaural).

Scene-based Audio is a 3D audio technology that utilizes Higher-Order Ambisonics (HOA). HOA enables precise capturing, efficient transmission, and immersive playback of 3D audio sound fields on various devices, including headphones, diverse loudspeaker setups, and soundbars [7].

1.2 Object-based audio

The main idea of object-based audio is to represent a given sound as an object. Such an object has a recorded audio signal and accompanying metadata. This metadata describes various characteristics of a given sound source, such as its location in three-dimensional space. Using the right processing, such an object can be placed at any point on a "virtual sphere" that reflects the real space in which the listener is located. Thanks to this procedure, the listener can feel the sensation of recognizing the direction from where the sound is coming. The entire sound content described by the emitted sound objects creates a virtual soundstage [8]. The big advantage of this approach is that (unlike the channel approach), such sound can be reproduced on any listening system.

1.3 Binaural sound, HRTF, and head tracking

An interesting technique is binaural audio, which differs from the typical stereo in that it considers aspects such as the Interaural Time Difference (ITD), i.e. head attenuation of certain frequencies, the effect of Interaural Level Difference (ILD), i.e. different times, that sound waves take to reach ears and other features that affect binaural listening. ILD and ITD cause a certain difference between the sounds reaching each ear separately. Binaural recordings can be done in various ways, for instance, using special microphones that are arranged in relation to each other as if they were to imitate human ears. Recordings made in this way take advantage of the natural ability of humans to precisely locate the sound source in the space around the listener [4].

HRTF (Head Related Transfer Function) set is a set of functions that determine how the ear perceives sound from a given point in space. They consist of numerous impulse responses measured at the entrance to the ear canal. They reflect the effect of structure, density, or head size on the perceived sound, which includes such phenomena as the amplification of some frequencies and attenuation of others. More specifically, they describe how the perceived sound (parameterized as frequency and source location) is filtered through the various components of the listener's body before reaching the inner ear [5]. HRTFs are used during binaural sound reproduction in headphones to produce the sensation of full sound perception.

Each person can get an individually measured set of HRTFs. To measure such an individual set, small measurement microphones are placed at the entrance to the ear canals of the listener. The listener is usually surrounded by a speaker array, which can be arranged in the plan of, for example, a sphere or a circle. If it is on a circular plan, the listener must rotate to collect impulse responses from each direction [9]. There is also a way to measure overall HRTFs. To do so, a combined setup of artificial head and body consisting of a torso, arms, and head that have microphones embedded in their ears to record a given audio signal is used [9].

A well-chosen set of HRTFs can give a truly qualitative impression of the soundstage. Moreover, the typical HRTF dataset is measured from an evenly distributed set of measurement points, which allows for another functionality that even further enables immersion in the sound scene. This feature, combined with dynamic tracking of the listener's head position, allows for exploration within the virtual sound scene. However, an additional sensor is required to take advantage of this feature – the so-called head tracker. This is a set of accelerometers and other sensors that can track the position (typically only the rotation) of the listener's head. Such a head tracker can be an external device that can be mounted on the top of the headphones. There are also other devices that enable dynamic tracking of head movements. These include e.g., Apple Air Pods [10], which use multiple gyroscopes as well as accelerometers for music listening, Oculus goggles [11], which employ external cameras, sensors, and 6 degrees of freedom in virtual reality applications (VR), and face-tracking cameras [12], based on image analysis algorithms, machine learning, or tracking of key facial points applied to facial recognition technology, video conferencing, or security systems.

2. The design of the experiment

The idea of the experiment was to combine the object-based audio with the scene-based audio so that the object signals provided localizability of sound sources and more direct perception, and the recorded soundstage more closely reflected the fidelity and veracity of the real scene. To obtain this type of material, a live recording was made. The recorded band consisted of piano, violin, electric bass, and alto soloist. Musicians were spaced in a semicircle, as presented in Figure 2. This arrangement enabled the utilization of the 360 space given by the immersive nature of the recording. The signals were collected with a higher-order ambisonic microphone array (namely, Zylia ZM-1) and spot microphones.

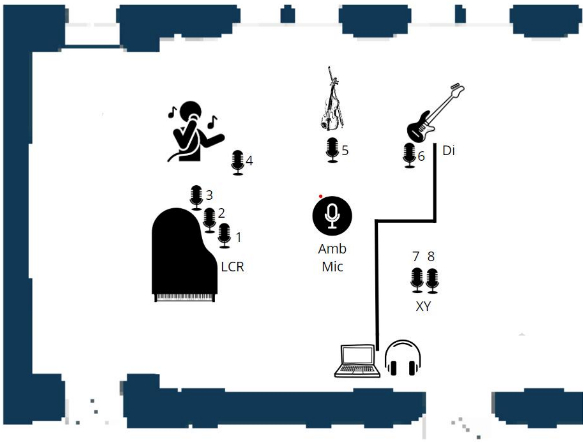


Figure 2. Recording setup

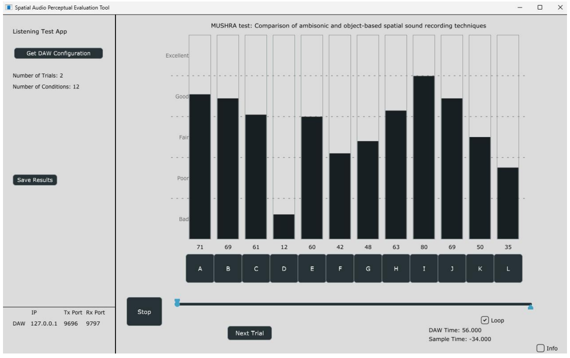
3. Post-production stage

One of the most important aspects was normalizing the individual recordings collected by the support microphones in relation to the ambisonic recording. To do this, the beamforming functionality of the microphone array was utilized, provided by the microphone's manufacturer. This solution, namely the Zylia Studio PRO plug-in [27], works with the native microphone's signals – as opposed to beamforming in the ambisonic domain. In this way, no intermediate rendering (to and from ambisonics) was performed. A set of narrow beams pointed to the musicians' positions was created and compared against the spot microphone signals. After this procedure, the volume of the sound from the object track was adjusted so that the differences in LUFS-S and LUFS-I were negligible (± 2 LUFS).

After normalization, the object tracks were encoded into ambisonics using the IEM MultiEncoder plug-in [13], according to the actual setting as it was during the recording. The ambisonic-encoded object-based signals were rendered to the 1st, 3rd, and 5th ambisonic orders. As to the scene-based ambisonic recording, it was already recorded with 3rd order of ambisonics; therefore, downmixing to 1st order was trivial. Additionally, the upscaling to the 5th ambisonic order was performed via the Imager by AudioBrewers [14]. In this way, both scene-based and object-based signals were prepared in the 1st, 3rd-, and 5th- ambisonic orders. Thus, 12 listening samples were created: a combination of each order of ambisonics with each order of converted objects (9 combinations) and 3 individual variants: 3rd-order ambisonics, 5th-order ambisonics, and 5th-order ambisonic-encoded objects. Such order mixing can result in various effects, such as incoherent auditory images or problems with spatial matching [15], but also a positive effect can occur, like the impression of externalization [17] or a bigger immersion sensation.

4. Auditory test stage

The auditory test was conducted using the Sapetool software [21], which allows for a MUSHRA-like test procedure with full trial randomization.



Spatial Audio Perceptual Evaluation Tool

Listening Test App

Number of Trials: 2

Number of Conditions: 12

MUSHRA test: Comparison of ambisonic and object-based spatial sound recording techniques

	IP	Tx Port	Rx Port
DAW	127.0.0.1	9696	9797

☑ Loop

DAW Time: 56.000

Sample Time: -34.000

☐ Info

Figure 3. Exemplary test session using the Sapetool interface [21]

A total of 21 listeners participated in the test. Each person had experience with professional audio to a significant degree, but with different professions: live sound realization, studio recording, audio processing algorithms and spatial sound. Participants were tasked with evaluating 12 samples on a scale from 0 to 100 in two trials: in the first trial, they were asked to rate the samples in terms of the degree of immersion and, in the second trial, in terms of the localizability, that is, the best ability to determine where the sound was coming from. Participants could listen to the samples in any order and as many times as they wished in each trial. The order of the samples was randomized each time (including between each trial).

The room in which the test took place had controlled acoustic adaptation conditions with short reverberation time. This experiment was performed using reference-grade open-back headphones with electrostatic membrane drivers (namely, Stax SR007 mkII) with a dedicated headphones amplifier (Stax SRM-727II). In addition, the head tracking device was used (namely, the Supperware head tracker 1 [22]), which provided dynamic, low-latency head tracking with a 100 Hz refresh rate. To provide dynamic sound reproduction and binaural listening, plugins from the IEM plug-in suite were employed: IEM SceneRotator [13], which allows communicating with head tracker via OSC, and the IEM BinauralRenderer [13] that uses the magnitude least squares method [17, 30] to convert ambisonic material into binaural signals directly. It is also worth noting that the binaural renderer uses HRTFs from the Neumann KU100 dummy head [18, 19, 29]). No far-field headphone equalization was used.

5. Results

The results from the immersion rating and localizability rating tasks are presented in Figs. 4 and 5, respectively (**figures on last page**). The denotations should be understood as follows: the letter corresponds to the recording technique (A – ambisonics, O – objects) and the number denotes the corresponding spatial resolution defined by the specific ambisonic order (for instance, A3 + O1 means 3rd-order ambisonics with 1st-order-encoded objects). The first 9 plots correspond to combined variants and the latter 3 plots correspond to individual variants. An omnibus Welsch's test was performed which showed statistical significance. Therefore, a post hoc single-factor ANOVA analysis (Tukey's analysis) was performed. The statistically significant differences are marked on charts with the corresponding labels.

5.2 Immersion ratings

Analyzing the chart with the distribution of immersion ratings (Figure 4), it can be concluded that this parameter received the lowest ratings for samples combined with 1 row of objects. In the combined samples where it occurred (1st- and 5th-order ambisonic), the average rating also increased as the row of objects increased. An interesting case is the set of combined samples for 3rd-order ambisonic, where the A3 + O3 variant had the highest rating. The highest rating in terms of immersion will be given to the sample in which the 5th-row objects alone were present.

5.3 Localizability ratings

Analyzing the chart with the distribution of localizability ratings (Figure 5), one can quickly notice a trend: as the order of objects increased, the ratings also increased. The order of ambisonics had less of an impact: while there is a slight increase in ratings with the increase in ambisonic order where objects in the 1st order are present, this trend did not persist with objects of the 3rd- and 5th- orders. The highest ratings from this set were received by A3 + O3 and A3 + O5 (not as expected: A5 + O3 and A5 + O5). Analyzing the non-combined samples, it is observed that A3 was rated similarly to A1 + O1 and A5 similarly to A3 + O1 (among the lower ratings). Meanwhile, O5 received the second-highest average rating. In this case, the highest rating (slightly higher than A5) was given to the sample A3 + O5.

6. Conclusions

The results obtained indicate that the best immersion variant was individual object stimulus rendered in 5th ambisonic order spatial resolution. In terms of localizability, the 3rd-order ambisonic combined with the 5th-order objects variant gave the best performance.

The object-based ambisonic-encoded signals, recorded at close microphone distances, exhibited low reverberation and a high level of proximity to the listener's position, making it, to some extent, counterintuitive that this stimulus was the most immersive. This unexpected outcome could be partially attributed to the so-called room divergence effect [25], as the experiment was conducted in a studio with a short reverberation time.

The distinct characteristics of the studio may have enhanced the perceived immersion of the audio despite or perhaps because of its clearer and more direct qualities. This phenomenon suggests that our traditional understanding of what makes audio immersive may need adjustment based on the specific acoustic properties of the recording environment.

Additionally, this was likely influenced by the fact that the headphones were open-back. Therefore, to ensure greater reliability of the results, the test could be repeated with the same group of people but in a different acoustical environment with a longer reverberation time.

Another aspect could be the fact that the spot microphones used during the recording were of very high quality. The combined quality of the spot microphones could outperform the tonal balance of the MEMS-based ambisonic microphone, which could be an important factor in the panel of audio engineering experts. For instance, this could have been one of the reasons why the O5 variant received such high ratings. In the subsequent investigation, using the same model of spot microphone for each application could be considered, ensuring their cumulative quality provides a similar level of quality as the ambisonic microphone.

References

[1] Peters, N., Sen, D., Kim, M.-Y., Wuebbolt, O., & Weiss, S. M. (2015, October). Scene-Based Audio Implemented with Higher Order Ambisonics (HOA). SMPTE 2015 Annual Technical Conference and Exhibition. IEEE. doi:10.5594/m001651

[2] M. Neukom, “Ambisonic Panning,” in AES 123rd Convention, New York, NY, USA, 2007

[3] Philip Coleman et al., “An Audio-Visual System for Object-Based Audio: From Recording to Listening”, IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 20, NO. 8, AUGUST 2018

[4] J. C. Middlebrooks, D. M. Green, “Sound localization by human listeners,” Annual Review of Psychology, 1991

[5] https://itigic.com/pl/hrtf-listening-to-positional-audio-in-games/ (06.10.2023)

[6] A. Rosiński, “Microphone Techniques in Stereo and Surrounding Recording,” Jagiellonian University Press, Kraków, 2022

[7] F. Olivieri, N. Peters, D. Sen, “Scene-Based Audio and Higher Order Ambisonics: A technology overview and application to Next-Generation Audio, VR and 360° Video”, EBU Operating Eurovision and Euroradio, technical review, 2019

[8] X. Sun, “Immersive audio, capture, transport, and rendering: a review,” industrial technology advances, 2021

[9] B. Mróz, B. Kostek, “ Externalization in binaural, ambisonic auralization of directional sources,” XXVIII Seminar on Applications of Computers to Science and Technology, Faculty of Electrotechnics and Automatics, Gdansk University of Technology, no. 60, 2018 (*in Polish*).

[10] Apple, https://support.apple.com/pl-pl/guide/airpods/dev00eb7e0a3/web (11.05.2024)

[11] Meta, https://www.meta.com/pl-pl/help/quest/articles/headsets-and-accessories/using-your-headset/turn-off-tracking/ (11.05.2024)

[12] E. Supriyanto, Y. Kee Jiar, T. Yong Oon, T. Meng Kuan, “Facial Tracking based Camera Motion Control System,” Progressive Health Care and Human Development Research Group Universiti Teknologi Malaysia UTM Skudai, 81310 Johor MALAYSIA

[13] IEM Plug-in Suite, Institute of Electronic Music and Acoustics, https://plugins.iem.at/ (11.05.2024)

[14] Ab Imager, AudioBrewers, https://www.audiobrewers.com/plugins/p/ab-imager (11.05.2024)

[15] B. Mróz, P. Odya, P. Danowski, M. Kabaciński, “A commonly-accessible toolchain for live streaming music events with higher-order ambisonic audio and 4k 360 vision”, in AES International Conference on Spatial and Immersive Audio, Huddersfield, UK, 2023

[16] G. Reardon et al., “Evaluation of Binaural Renderers: Externalization, Front/Back and Up/Down Confusions,” in AES 144th Convention, Milan, Italy, 2018

[17] Zotter, F., & Frank, M. (2019). Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality. Springer International Publishing. doi:10.1007/978-3-030-17207-7, chapter 4.11.2, page 89

[18] Neumann, “KU 100 Dummy Head Product Information”, ku100.pdf

[19] Sofa HRTFs database, http://sofacoustics.org/data/database/thk (11.05.2024)

[20] Neumann KU 100 image, https://www.thomann.de/pl/neumann_ku100.htm

[21] T. Rudzki, “REAPER-MUSHRA: Listening test tool for DAW-based perceptual evaluation of spatial audio”, https://github.com/trsonic/sapetool/tree/tr-develop (29.04.2024)

[22] Supperware Head Tracker 1, https://supperware.co.uk/headtracker-overview

(09.05.2024)

[23] S. Bech, N. Zacharov, “Perceptual Audio Evaluation. Theory, Method and Application”,

John Wiley and Sons Ltd, Copyright 2006, chapter 9

[24] R. Tanious, R. Manalov, “Violin Plots as Visual Tools in the Meta-Analysis of Single-Case Experimental Designs,” Methodology journal, 2022, Vol. 18(3), pages 221-238

[25] S. Werner, F. Klein, T. Mayenfels, K. Brandenburg, "A Summary on Acoustic Room Divergence and its Effect on Externalization of Auditory Events," Electronic Media Technology Group Technische Universität Ilmenau, Ilmenau, Germany

[26] A. Farina, “Performing not linear processing (De-noising, Compression, Limiter, etc.) on High Order Ambisonics signals using Adobe Audition CC and the SPS approach,” http://pcfarina.eng.unipr.it/Aurora/Ambisonics-Denoising.htm (11.05.2024)

[27] Zylia, https://www.zylia.co/zylia-studio-pro.html (11.05.2024)

[28] J. Peterson, H. Lee, “3D audio”, copyright 2022

[29] Bernschütz, Benjamin. "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100", Proceedings of the 40th Italian (AIA) Annual Conference on Acoustics and the 39th German Annual Conference on Acoustics (DAGA) Conference on Acoustics. 2013. http://audiogroup.web.th-koeln.de/ku100hrir.html (11.05.2024)

[30] Schoerkhuber, Christian; Zaunschirm, Markus; Hoeldrich, Robert. "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," Fortschritte der Akustik, DAGA, 2018

Ratings charts:

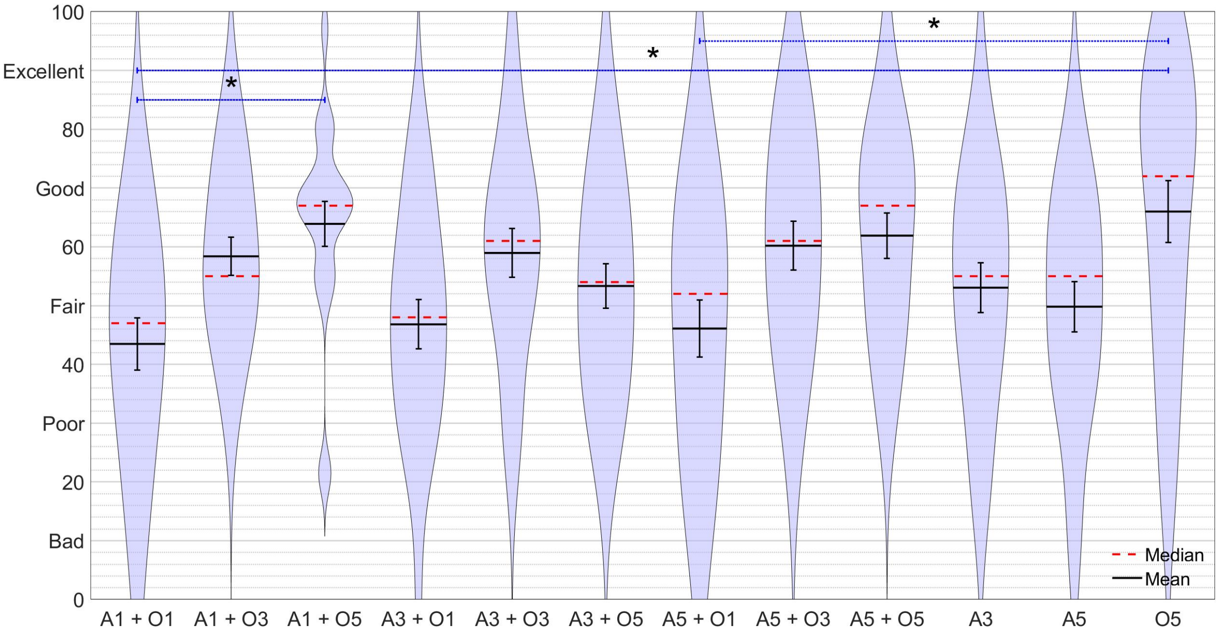


Figure 4. Distribution of immersion ratings with statistically significant comparisons from ANOVA test (“*” symbols represent p < 0.05). The mean value is marked with a solid line, the median with a dashed line. Confidence intervals marked as vertical lines. The text labels on the vertical axis reflect the ratings ranges from the Sapetool interface [21].

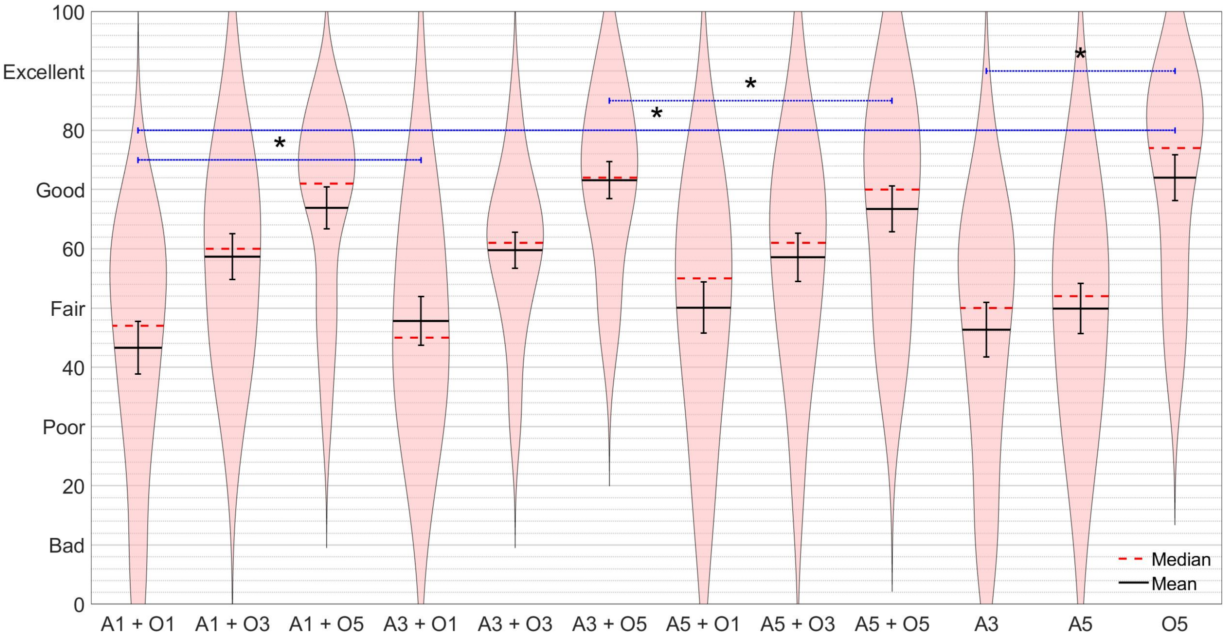


Figure 5. Distribution of localizability ratings with statistically significant comparisons from ANOVA test (“*” symbols represent p < 0.001). The mean value is marked with a solid line, the median with a dashed line. Confidence intervals marked as vertical lines. The text labels on the vertical axis reflect the ratings ranges from the Sapetool interface [21].