Audio Engineering Society

Convention Express Paper 407

Presented at the AES 160th Convention
2026 May 28–30, Copenhagen, Denmark

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Perceptual Evaluation of the MPEG-I Immersive Audio Standard

Andreas Silzle[1], Pablo M. Delgado[1], Alejandro Restrepo Garcia[1], Sascha Disch[1], Leon Terentiv[2], Erlendur Karlsson[3], Jussi Leppanen[4], Ehara Hiroyuki[5], and Sam Jelfs[6]

[1]*Fraunhofer IIS, Erlangen, Germany*
[2]*Dolby Germany GmbH*
[3]*Ericsson AB*
[4]*Nokia Technologies*
[5]*Panasonic Holdings Corporation*
[6]*Koninklijke Philips N.V.*

Correspondence should be addressed to (andreas.silzle@iis.fraunhofer.de)

ABSTRACT

The recently finalized ISO international standard (IS) on MPEG-I Immersive Audio enables interactive six-degrees-of-freedom (6DoF) audio rendering for a multitude of virtual-reality and augmented-reality (VR/AR) acoustic scenarios and applications. It consists of comprehensive modeling of room acoustics and intricate acoustic phenomena, including e.g. occlusion, reflection, transmission and diffraction caused by sound obstacles, Doppler effect, and dynamic environment changes triggered by user interactivity. This paper describes concept, methodology and results of the final verification test of the standard. The perceptual quality of the renderer was assessed, stabilized by perceptual anchors, in an interactive listening test using different in- and outdoor acoustic scenes. The considerations behind the selection of the test setup and the design of suitable anchors will be discussed. In the test, 56 subjects participated distributed across six labs using the ITU-R Rec. BS.2132 multi-stimulus method on a 100-point scale for three conditions (IS, mid- and low-anchor) in 10 VR scenes plus two repetitions. The finally selected mid and low anchors have demonstrated stable quality across diverse scenes with progressive timbre and spatial degradations, and the results show a clear separation of the conditions (IS > mid > low); the low anchor was stable (around 16 points median value) while the mid anchor varied by scene (around 47 points). Absolute ratings are scene dependent and their quartile range for some scenes can exhibit up to 20 points. The median value obtained for the IS varied slightly per lab, indicating that apparently some labs had more critical listeners than others. The IS is rated with a median of 84 points among all labs, which is the "excellent" region of the scale.

1 Introduction

The new MPEG-I Immersive Audio standard [1, 2] enables high-quality storage, metadata transmission and audio rendering of 6DoF sound in complex virtual environments. It supports different classes of audio sources, ranging from point sources to sources with spatial extent. Moreover, it includes the modeling of specific radiation patterns, such as those stemming from musical instruments. Based on a geometric description of acoustically relevant scene details (e.g., walls, doors, moving objects) with their sound-related properties (e.g., reflecting, occluding and transmission coefficients), the rendering process incorporates comprehensive modeling of room acoustics and complex acoustic phenomena, specifically accounting for occlusion, reflection, and diffraction caused by solid sound obstacles. Furthermore, the standard accommodates dynamic effects (e.g. Doppler) and facilitates real-time environment changes (triggered by both user interactivity and content creator instructions) [3].

The paper describes concept and methodology to subjectively assess the rendering sound quality of this standard. The results of the final verification test are presented and discussed. Among the test subjects, around half of them were MPEG-I experts and the other half were expert listeners that were unfamiliar with MPEG-I rendering. The test results are presented for both individual groups as well as for the aggregate cohort.

2 Background

A meaningful assessment of 6DoF interactive audio is a unique challenge [4], since there is no ground truth available for direct comparison, except for the listeners' *inner reference* their subjective expectation of plausible natural acoustics [5]. The absence of an easily defined external reference, that has existed in all former ISO/MPEG audio codec evaluations, prohibited the use of the standard methodology ITU-R Rec. BS.1534 [6] using a MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA). Any evaluation of 6DoF interactive audio is further influenced by the test scene design, its size and acoustic complexity, the used audio signals, the interactivity features, and the experience of the subjects with 6DoF environments in general and the specific test scenes in particular (e.g. pre-knowledge of critical listening positions, paths, movements and interactions with the scene environment).

Different so-called Quality Features (QF) in the perceptual domain contribute to the Overall Quality for VR applications. The application-dependent contributions of the QFs to the Overall Quality is evaluated by expert surveys and their perceptual quality is evaluated by listening experiments, [7]. Relevant QFs are listed in the following:

- *Location Accuracy*
- *Timbre*
- *Loudness Balance*
- *Auditory Spaciousness*
- *Reverberance*
- *Dynamic Accuracy*
- *Absence of Artefacts*

In the perceptual evaluation of MPEG-I Immersive Audio, Overall Quality was regarded as the most important and appropriate measure. Because the standard is rather general and not tied to any specific application scenario, an absolute grading of individual QFs is not quite reasonable, [7].

3 Test Methodology

The verification test for the MPEG-I Immersive Audio standard was conducted across six international laboratories (Dolby, Ericsson, Fraunhofer IIS, Nokia, Panasonic, and Philips) and involved 56 expert listeners (after post-screening). The test was performed within a real-time 6DoF VR environment where subjects could navigate freely through 10 diverse virtual scenes to assess the Overall Quality and Plausibility of the audio, which should be evaluated regarding the internal reference of the subjects. The setup utilized a multi-stimulus method following ITU-R Rec. BS.2132 [8], comparing the International Standard (IS) renderer against mid- and low-quality anchors designed to induce progressive timbre and spatial degradations. Within the VR environment, subjects interacted with a semi-transparent GUI and used a virtual laser pointer to switch seamlessly between time-aligned rendering conditions for direct comparison (see Figure 1). The general structure of the real-time VR listening test software, the Audio Evaluation Platform (AEP) [4], is illustrated in Figure 2.

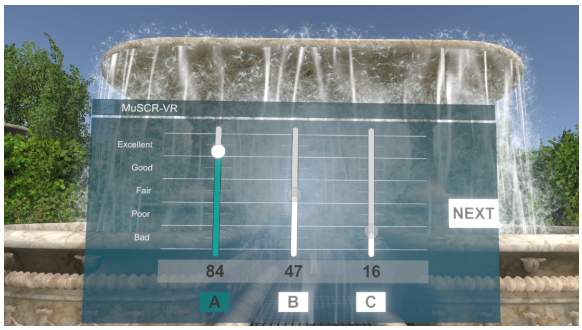


Fig. 1: Graphical user interface (GUI) of the Multi-Stimulus Comparison Rating VR test (MuSCR-VR). It can be evoked and freely positioned by the subject during the test. Conditions can be switched and scores be set using a virtual laser pointer.

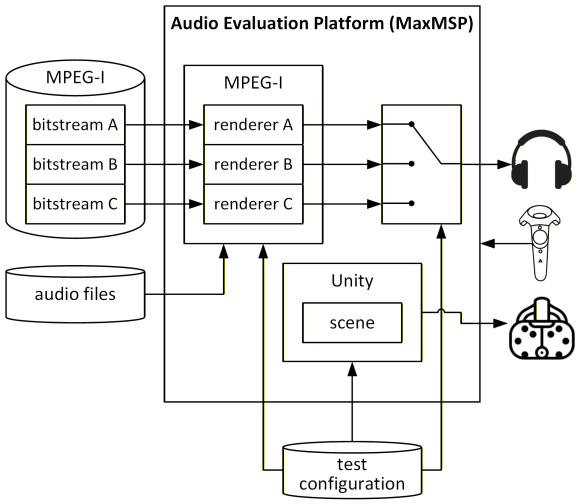


Fig. 2: Audio Evaluation Platform (AEP) for realtime testing of different 6DoF renderers.

3.1 Test Method

After extensive investigations on a suitable method for subjective assessment in 6DoF [4], a multi-stimulus test based on ITU-R Rec. BS.2132 [8] , without an external reference was chosen. This method was adapted to be used in VR and is hereafter referred to Multi-Stimulus Comparison Rating VR (MuSCR-VR) test.

The subjects were provided with detailed instructions on the test. The instructions made clear that, due to the absence of a hidden reference, no mandatory 100 points (maximal score) for one of the conditions were required, unlike demanded in traditional ITU-R Rec. BS.1534 MUSHRA [6] tests, used often in ISO/MPEG audio tests until now. As verbal anchors for the 100-point scale, the standard five terms to characterize successive 20-point ranges are used, in order of decreasing quality: "*excellent*", "*good*", "*fair*", "*poor*" and "*bad*".

Rather than being an end-user test of a specific consumer application, the verification test was meant to be a most critical evaluation of the technology itself. Therefore, subjects were required to be experts in audio evaluation, as it was the case in previous ISO/MPEG audio tests. The subjects were also categorized in either general audio evaluation experts or MPEG-I listening experts (developers of the standard and well experienced with 6DoF interactive rendering). To ensure reliable listening test data, two scenes were repeated at the end of the sequence to verify subject consistency.

The details of the verification test are discussed in the following sections.

3.2 Listener Instructions

Proper listener instructions are very important for obtaining consistent results that are truly representative for the technology under test, the more so in the absence of an explicit reference condition. The instructions focused on the plausibility and realism of the audio embedded in the interactive audio-visual scene and advised subjects to balance their grading between critical listening and overall listening impressions. The following instructions were given to the subjects:

"You will move freely in several virtual 3D environments; your task is to rate the quality of the audio reproduction with a single Overall Quality score (0 bad - 100 excellent). This rating should be an average of the impressions collected from critical and overall listening approaches, explained below."

Since the *critical listening* aspect was supposed to pinpoint possible renderer shortcomings, additional instructions were provided to the test subject for further guidance:

"Listen for:

- *Breaks in continuity when sound transitions between occluded/un-occluded*

- *Missing or unnatural diffraction around edges*
- *Inaccurate early reflections from irregular surfaces*
- *Unrealistic reverberation or leakage between rooms*
- *Scoring: mark low scores where you detect artifacts or implausible behaviors*
- *Features specific to each scene (mentioned in the pre-room of each scene)"*

The *overall listening* aspect targeted a result of an Overall Quality and Plausibility rating and was described by the following sentence:

"Listen more holistically: is the scene plausible? Avoid focusing solely on the flaws you found in the critical pass."

3.3 Anchor selection

In subjective audio quality assessment for VR/AR, *anchors* are deliberately degraded rendering processes included in the listening test to stabilize the rating scale and provide interpretable benchmark quality levels. The MPEG-I Immersive Audio verification test anchors were selected by the MPEG audio working group to balance scientific rigor—i.e., producing stable rating scales and consistent, reliable results [6]—with industry relevance, i.e., providing benchmark results that compare the system against understandable spatial audio degradations representative of VR/AR applications rather than generic degradations [9]. The agreed-upon requirements demanded that there must be a low-quality anchor for providing proper scale usage, and there should be a mid-quality anchor that needs to render perceptually stable audio quality across diverse scenes.

It was interesting to see in a pilot test that the sole timbral degradation of the usual 3.5 kHz low-pass filter resulted in rather high scores with a median value of 48 points and, depending on the scene, a sometimes surprisingly high individual score. This suggested that an additional degradation in the spatial perceptual domain would be required for a suitable anchor.

Both anchors need to employ consistent, explainable degradation processes, both in timbral and spatial perception domains. The selection process involved extensive pilot testing of various renderer modifications, including a basic renderer configuration without modeling early reflections, diffraction, extent and portals. However, the performance of such a basic renderer was found to be too inconsistent across the test item set, since the perceptual importance of these dedicated acoustic effects is largely scene-dependent. Further investigations compared modified HRTFs, specifically evaluating Interaural Level Difference (ILD)-only models against stereo panning laws; the audio working group agreed that panning offered more stable and explainable spatial degradation, easily understood by most audio experts.

The final anchor selection resulted in the following two anchor conditions:

- *Low anchor*, implemented as a mono mixdown with a 3.5 kHz low-pass filter to represent minimum audio reproduction quality
- *Mid anchor*, implemented using stereo amplitude panning with a 7 kHz low-pass filter. The stereo amplitude panning employed the Linear Constant Gain Sum amplitude panning law, with additional saturation of the channel gains to prevent them from falling to zero. The gains only varied with azimuth and ignored the elevation. The gains are illustrated in Fig. 3.

Only with these selected mid- and low-anchors consistent quality rating across diverse scenes with progressive timbre and spatial degradations could be achieved (see Subsection 4.5).

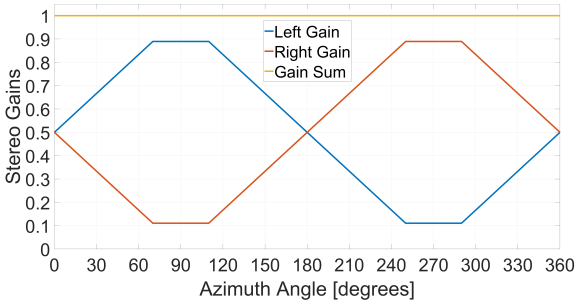


Fig. 3: Stereo amplitude panning gains using the linear constant gain sum law with saturation.

3.4 Practical Considerations

Several other external technical factors influenced the design of the testing methodology concept.

For the verification test, the official MPEG reference software implementation was used [10]. The reference software was running on contemporary commercial VR hardware (see [3]). However, a multi-stimulus test with instant condition switching requires as many renderers as existing conditions to run simultaneously on the test machine. This puts a limit on the maximum number of conditions that can be assessed on a given machine without overloading it.

Furthermore, for both technical and conceptual reasons, a comparison with a commercial benchmark renderer was not feasible. Such a comparison would require porting all test scenes to the scene-description formats of other auralization engines. Besides technical incompatibilities, this would involve extensive parameter tuning, making the validity of the comparison questionable. In addition, there might be legal constraints on publishing results for commercial products without explicit consent.

A comparison with a real environment is impractical for several reasons. It would restrict the test to a single simple, relatively small indoor space and would not allow scenarios such as a large, highly reverberant room (as in the cathedral simulation) or outdoor scenes with extended sound sources (such as the waterfall or the large park fountain that can be explored from far, near and even inside). Moreover, such a real-world setup could not be replicated across multiple laboratories participating in the test, and would therefore not allow investigation of lab-dependent effects.

Most importantly, the design goal of the MPEG-I Immersive Audio renderer is plausibility rather than strict authenticity. For most anticipated applications, this level of realism is sufficient. Additionally, truly authentic aural reproduction is unlikely to be achievable with the capabilities of current and near-future integrated circuits in head-mounted displays.

3.5 Test Items and Conditions

The VR scenes that were selected for the verification test set out of the pool of available MPEG-I test scenes are described in Table 1. Some of these scenes were new or modified scenes, which were not used in the development of the standard.

The test set was assembled with the focus on:

- Offering a wide variety of scenes settings (in- and outdoor in very different environments).
- Covering all types of sound sources, including dynamic sources and multiple types of occluding/reflecting/diffracting geometries.
- Challenging multiple stages of the renderer with sufficiently complex scenes.
- Balancing critical scenes (known to challenge the renderer) with representative scenes (permitting a fair assessment of overall audio quality).
- Supporting assessment of plausibility over hyperrealism by avoiding any bias to compelling yet unnatural sound (since no explicit reference is available).
- Ensuring multi-modal coherence of visuals, motion and sound by realistic scene authoring.
- Providing interactivity (switchable and movable sound sources, triggered sliding doors).
- Limiting the number of scenes to bound the necessary test time.

To level align the different scenes, some level adjustments were made to enable all testing at the level set in the familiarization scene.

In the test, three conditions were rated: MPEG-I Immersive Audio international standard (IS), and the Mid and Low anchors described in Subsection 3.3.

3.6 Familiarization and Level Adjustment

Each test run began with a dedicated familiarization scene. This scene was first be used for adjusting the headphone playback level, with the subject remaining at the initial position. After this adjustment, the headphone level had to remain fixed for all subsequent scenes. The only exception was when a subject chose to walk inside the fountain in the Park scene, in which case the operator could temporarily reduce the headphone level. The familiarization scene also served to train the non-MPEG-I experts in using the AEP and its interaction controls. The familiarization scene was not scored and not used later in the test. The following test scenes were presented in random order.

4 Results

4.1 Test Subjects and Consistency

Two different expert groups evaluated the MPEG-I Immersive Audio renderer: firstly, the group of MPEG-I

Table 1: Test Scenes.

Name	Description
Canyon	Outdoor scene with reflections and an interactively movable boombox.
Cathedral	Indoor scene with high reverberation in two coupled rooms plus outdoor area.
Hospital	Many switchable sound sources, one moving sound source, three different rooms, interactive sliding doors.
Park	Extended sound source of a fountain and occluders like pillars, walls, and hedges.
Recreation	Four different in- and outdoors areas: swimming hall with high reverberation, waterfall as an extended sound source, piano bar, and outdoor rock concert.
Singer in the Lab	Singer in acoustically treated sound lab with 3D loudspeaker setup playing music.
Virtual Basketball	Features voxel-based geometry; gym hall with interactive ball animation.
Downtown Drummer	Drummer in outdoor street scene with an occluding bus.
Multizone Music	Multi-zone HOA simulation of many loudspeakers in a room.
Garage	Features granular synthesis; car circles in large parking garage with occluding wall; listener can move around freely and evaluate car sounds and garage acoustics.

developers, who are very familiar with the algorithm and all the test scenes. Secondly, a group of audio expert listeners from other fields such as sound engineering, automotive audio tuning, game development and others. The number of expert listeners from the different groups and labs is presented in Table 2, after post-screening (details below).

Table 2: Number of subjects after post-screening.

Lab	# MPEG-I experts	# Other experts	# All listeners
Dolby	4	4	8
Ericsson	4	5	9
Fraunhofer IIS	10	4	14
Nokia	2	8	10
Panasonic	6	0	6
Philips	4	5	9
Sum	30	26	56

At the end of each test, two scenes (Park and Cathedral) were repeated to assess subject rating consistency. For each of the three conditions, the difference between the scores for the first and second presentation was calculated. The distributions of these differences (before post-screening) were then analyzed, see Table 4. Outliers with an absolute difference greater than 30 points in at least one of the conditions for either of the repeated scenes was defined as the threshold for post-screening. Based on this criterion, five outlying conditions from four subjects were identified, and all results from these four subjects were removed. These test subjects were given the opportunity to repeat the test, and one subject did so successfully. Thus, data from 56 subjects were included in the final evaluation.

Table 3: Consistency results of 21 repetitions.

	meanDiff	CI low	CI high
Cathedral	0.2	-2.7	2.8
Park	3.7	0.8	7.4

4.2 Test Duration

The average test duration for the MPEG-I experts, the other experts and all are presented in Table 4.

Table 4: Average test duration for expert groups.

	MPEG-I experts [minutes]	Other experts [minutes]	All [minutes]
Mean	35.8	37.0	37.8
Std.dev	16.6	13.3	14.4

On average the test duration was 38 minutes per subject, with a noticeable difference between the labs. The total test time over all labs with the 59 subjects was 37 hours.

4.3 Statistical Analysis

Table 5 shows the results of an ANOVA across all data for multiple factors. As can be seen, all factors are significant, with *Condition* being the factor that explains most of the variance in the data.

Table 6 shows the results of a Tukey HSD (Honestly Significant Difference) pairwise comparison over conditions, showing that all conditions are significantly different from each other. Table 7 shows the Tukey HSD comparisons for the labs, where only the scores from Fraunhofer IIS and Nokia are shown to have significant differences, all other pairwise-comparisons are not significantly different. Given these results, it would appear fair to average across scenes and labs for the final analysis.

Table 5: Results of the ANOVA over all data.

	Sum of Squares	df	F	p
Condition	1212260	2	3565.7	< 0.001
Scene	5636	9	3.7	< 0.001
Lab	8132	5	9.6	< 0.001
Lab×Cond.	9202	10	5.4	< 0.001
Scene×Cond.	9038	18	3.0	< 0.001
Residual	277933	1635		

Table 6: Tukey HSD – pairwise comparison by condition.

Cond. 1	Cond. 2	Mean Diff.	p-adj	lower	upper
IS	Low A.	−65.8	0.0	−67.7	−63.9
IS	Mid A.	−34.2	0.0	−36.2	−32.3
Low A.	Mid A.	31.5	0.0	29.6	33.4
Note: All three comparisons reject H_0 (same means)					

4.4 Combined Test Results of all Labs

Fig. 4 presents the pooled result scores of all labs, with 56 subjects, after post-screening, for the three tested conditions. The left side presents the results with mean and 95% confidence intervals, as often used in MPEG audio listening test presentations. The right side shows the data as boxplots with median, quartile and whiskers. The box plots provide a clearer representation of the distribution of the ratings.

4.5 Detailed Test Results

The further results are also presented as box plots, with median, quartile and whiskers. The non-parametric presentation does not require normal distribution of the results and is preferred for results from a small number of subjects per lab or scene. The whiskers are lines above and below each box covering 99.3% of the data.

Table 7: Tukey HSD – pairwise comparison by Lab (Do=Dolby; Er=Ericsson; FhG=Fraunhofer IIS; No=Nokia; Pa=Panasonic; Ph=Philips).

Lab 1	Lab 2	Mean Diff.	p-adj	lower	upper
Do	Er	0.24	1.00	−7.37	7.85
Do	FhG	−3.58	0.68	−10.53	3.36
Do	No	3.00	0.86	−4.43	10.43
Do	Pa	−0.23	1.00	−8.70	8.23
Do	Ph	0.64	1.00	−6.97	8.25
Er	FhG	−3.82	0.58	−10.52	2.87
Er	No	2.76	0.88	−4.44	9.96
Er	Pa	−0.47	1.00	−8.73	7.78
Er	Ph	0.40	1.00	−6.98	7.78
FhG	No	6.58	0.04	0.10	13.07
FhG	Pa	3.35	0.81	−4.29	10.99
FhG	Ph	4.22	0.47	−2.47	10.92
No	Pa	−3.24	0.86	−11.33	4.85
No	Ph	−2.36	0.94	−9.56	4.84
Pa	Ph	0.87	1.00	−7.38	9.13
Note: Only the comparison FhG vs. Nokia rejects H_0 (same means), all others fulfill H_0					

Fig. 5 presents the results of the three conditions per lab and the total results. The listening test results show a clear separation of the conditions (IS $>$ mid $>$ low). The low anchor was stable (around 16 points median value) and the mid anchor varied by scene (around 47 points). The IS condition is rated around 84 points among all labs, which is the "excellent" region of the scale. The median value for the IS of the different labs varied, some are a bit more critical than others.

In Fig. 6, the scene results of the pooled labs are presented. Two main observations can be made: the individual scenes are rated quite differently and the quartile range for some scenes can reach more than 20 points.

4.6 Different Expert Groups

Fig. 7 shows a rating comparison of the two expert groups per lab and Fig. 8 shows the same per scene. The quartile ranges of the other expert listeners were for some scenes, labs or conditions higher as for the MPEG-I experts. This tendency can also be observed in the "total" results per lab or per scene.

Comparison of the distribution of differences in the repetitions (after post-screening) for MPEG-I experts only and the other expert listeners only, showed that the other expert listeners have a slightly greater variance in their repeatability than the MPEG-I experts.

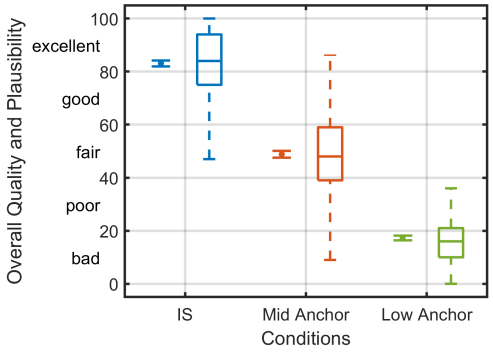


Fig. 4: Overall results for the three conditions, all labs pooled, (left) mean and 95% confidence intervals, (right) boxplot with median, quartile and whiskers, *#listeners* = 56, *#answers* = 560.

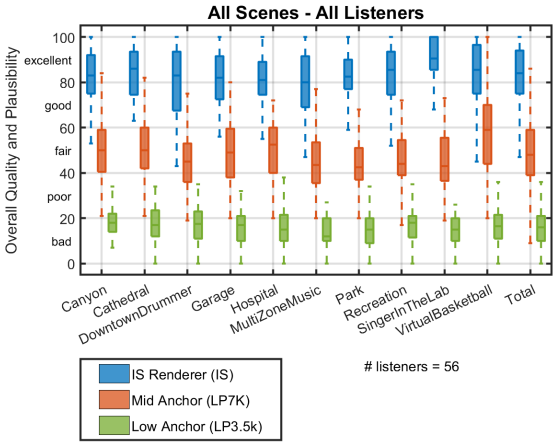


Fig. 6: Scenes: all labs and total, $N = 56$.

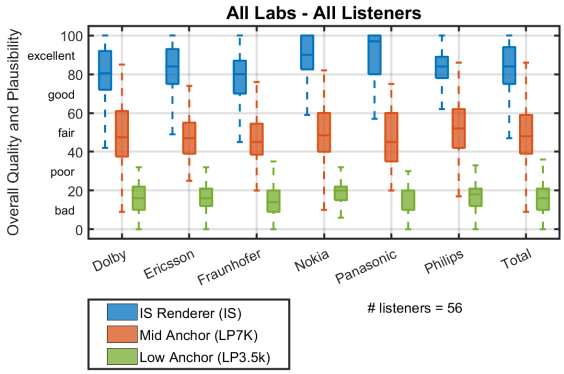


Fig. 5: Labs: conditions and total, $N = 56$.

4.7 Summary of Results

In summary, the listening test data shows that the IS is rated with a median of 84 points among all labs, which is the "excellent" region of the scale. In assessments utilizing only an internal reference, the scores usually do not go up to full scale (see in [11] Figure 12). However, this effect is nearly not visible in this investigation. The low-quality anchor remained largely stable across scenes, with a median of 16 points whereas the mid-quality anchor showed a larger variation around a median of 47 points. All selected scenes exhibit a clear and significant quality separation between the IS renderer, the mid- and the low-quality anchor. The subjects' repeatability shows quite good consistency for the two repeated scenes. Only 4 out of 59 subjects were removed because their absolute repetition difference was greater than 30 points in at least one of the three presented conditions. The ratings showed some scene dependency and the quartile range of the IS of many scenes is around 15–20 points. The median values for the IS of the different labs varied, some are a bit more critical than others. The group sizes of the MPEG-I experts and the other listening test experts were similar, 26 to 30. But the quartile range for all three conditions, per lab or per scene, and the repeatability is a bit larger for the non-MPEG-I experts. This repeated the experience of many former ISO/MPEG audio listening tests that the developers of the system are more critical listeners.

5 Conclusion

This paper describes concept, methodology and results of the final verification test of the MPEG-I Immersive Audio standard. In the MPEG-I Immersive Audio verification test, the overall perceptual quality and plausibility of the standardized renderer have been assessed by expert listeners. Grounded on two anchor conditions, a lower anchor satisfying minimum requirements for reasonable audio reproduction, and a mid anchor simulating a basic 6DoF audio rendering using well-known stereo amplitude panning, subjects rated the MPEG-I Immersive Audio condition overall as "excellent", confirming that the initial development goal of achieving high-quality interactive audio rendering has been accomplished.

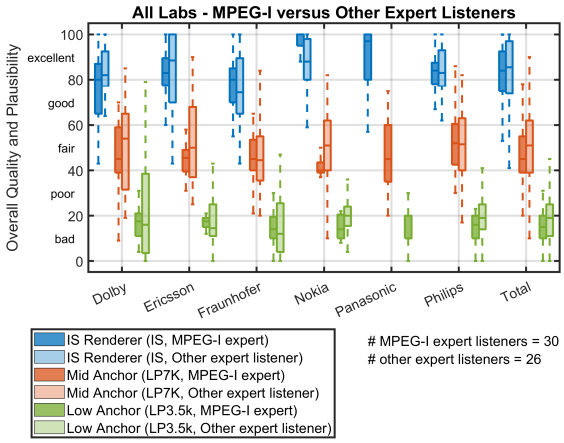


Fig. 7: Labs: MPEG-I experts versus other experts.

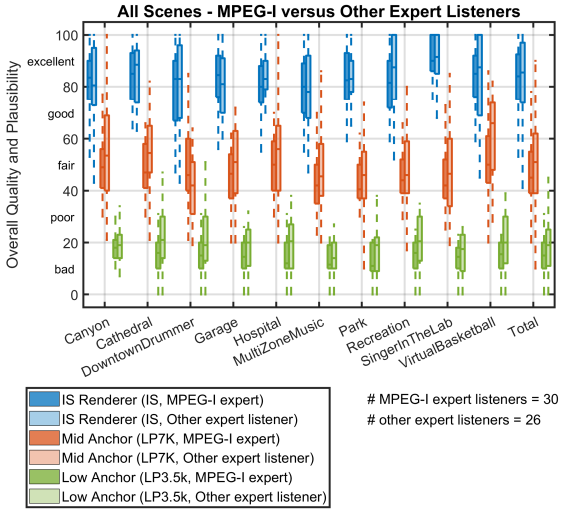


Fig. 8: Scenes: MPEG-I experts versus other experts.

Acknowledgements: We thank all who contributed to the definition of the verification test and its execution. Our sincere gratitude also goes to the listening test subjects.

References

[1] "ISO/IEC, 23090-4, Information technology — Coded representation of immersive media, Part 4: MPEG-I immersive audio," `https://www.iso.org/standard/84711.html`, 2025.

[2] Disch, S., Terentiv, L., Koppens, J., Falk, T., Leppänen, J., Munoz, I., Herre, J., Adami, A., Silzle, A., Fischer, D., Setiawan, P., Fersch, C., Hanschke, J.-H., Jelfs, S., Bruijn, W. d., Eronen, A., Mate, S., Dick, S., Borss, C., Delgado, P. M., and Trinidad, M. M., "MPEG-I Immersive Audio - The ISO/MPEG Standard for Virtual / Augmented Reality Audio," *Journal of the Audio Engineering Society*, (accepted for publication), 2026.

[3] Disch, S., Terentiv, L., Koppens, J., Falk, T., Leppänen, J., and Munoz, I., "MPEG-I Immersive Audio – The Technology of the New Standard for Virtual / Augmented Reality Audio," in *159th AES Convention*, Long Beach, USA, 2025.

[4] Herre, J. and Disch, S., "MPEG-I Immersive Audio – Reference Model for the New Virtual / Augmented Reality Audio Standard," *Journal of the Audio Engineering Society*, 71(5), pp. 229–240, 2023, doi: 10.17743/jaes.2022.0074.

[5] Brinkmann, F. and Weinzierl, S., *Audio Quality Assessment for Virtual Reality*, pp. 145–178, Springer International Publishing, Cham, 2023, ISBN 978-3-031-04021-4, doi:10.1007/978-3-031-04021-4_5.

[6] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems (MUSHRA)," Recommendation BS.1534-3, International Telecommunication Union, Geneva, Switzerland, 2015.

[7] Silzle, A., *Generation of Quality Taxonomies for Auditory Virtual Environments by Means of Systematic Expert Surveys*, Ph.D. thesis, Institut für Kommunikationsakustik, Ruhr-Universität Bochum, 2007.

[8] ITU-R, "Method for the Subjective Quality Assessment of Audible Differences of Sound Systems Using Multiple Stimuli Without a Given Reference," Recommendation BS.2132-0, International Telecommunication Union, Geneva, Switzerland, 2019.

[9] Rummukainen, O., Schlecht, S., Plinge, A., and Habets, E. A. P., "Evaluating Binaural Reproduction Systems from Behavioral Patterns in a Virtual Reality—A Case Study with Impaired Binaural Cues and Tracking Latency," *Journal of the Audio Engineering Society*, (9895), 2017.

[10] "ISO/IEC, 23090-34, Information technology — Coded representation of immersive media, Part 34: Immersive audio reference software," `https://www.iso.org/standard/89037.html`, 2025.

[11] Silzle, A., George, S., Habets, E., and Bachmann, T., "Investigation on the quality of 3D sound reproduction," in *Proc. of the International Conference on Spatial Audio*, pp. 334–341, Verband Deutscher Tonmeister e.V., Detmold, Germany, 2011.