Quality Testing for AR and VR in MPEG-I Immersive Audio

Thomas Sporer[1], Sam Jelfs[2], Jürgen Herre[3]
[1] Fraunhofer IDMT, 98693 Ilmenau, E-Mail: Thomas.Sporer@idmt.fraunhofer.de
[2] Philips Research, 5656AE Eindhoven, E-Mail: sam.jelfs@philips.com
[3] International Audio Laboratories Erlangen, 91058 Erlangen, E-Mail: juergen.herre@audiolabs-erlangen.de

1 Introduction

Traditionally, quality assessment of high-quality audio coding schemes is based on listening tests comparing the output of a coding system with its input. This input is named open reference. In Augmented and Virtual Reality (AR/VR), acoustic scenes are often not recorded but produced: They only exist as a scene description, and therefore no reference sound exists. MPEG Audio currently works on coding and rendering of 6DoF audio. In the course of the call for proposals (CfP) for MPEG-I immersive audio [1] it was necessary to select or develop a test environment and a test method which is adequate to select the core technology for the upcoming standard. For AR application, acoustical properties of the reproduction room are an additional input to the renderer enabling the adaptation of the rendered acoustic scene to the real room.

This paper describes the requirements for quality assessment, some already standardized test methods, and the way towards three candidate methods. From these methods, one was selected for the CfP, but all of them are foreseen to be used during the core experiment phase (i.e. improving the core technology by adding or replacing components) and the final verification test.

2 MPEG-I Immersive Audio

The MPEG audio group ISO/IEC JTC1 SC29 WG-6 is currently working on MPEG-I immersive audio both for VR and AR. The Call for Proposals (CfP) for this upcoming standard was issued in May 2021, and listening tests were conducted in December 2021. Winners were selected and a merge of the winners into one "reference model" (RM0) and "working draft" (WD0) happened. In the current "Core Experiment" (CE) phase additional technologies targeting either to improve RMx or adding features to it are proposed and evaluated. The finalization of the standard is expected for end of 2024.
On the way towards the CfP and the evaluation, several challenges and questions had to be faced:

- What is an appropriate content format as input?
- How to get access to 6DoF audio and video scenes for testing all acoustically relevant properties?
- How to set up a test environment to be used for selecting a winner in the CfP?
- Which test method should be used to evaluate competing proposals?

The first question has two parts: How to encode the audio waveforms ("essence") and how to specify the scene format including acoustic properties of the environment. For the audio essence, MPEG-H 3D Audio was selected as a basis. For the specification of acoustic properties of objects and rooms, the MPEG audio group defined metadata in the so-called "Encoder Input Format" (EIF) [6]. To adapt content to the acoustics of different reproduction rooms, the "immersive audio augmented reality Listener Space Description Format" (LSDF) [7] was specified.

The issue of test content was more difficult: The upcoming standard supports a lot of acoustical features in 6DoF which no other format could provide. Examples are sophisticated descriptions of audio sources (objects with size and directivity) and the acoustics of the enclosing room, partial and complete occlusion incl. dynamic occlusion like opening and closing doors, coupling of different rooms, and finally a natural Doppler shift for moving objects. Several of the partners involved in MPEG audio created test scenes using the EIF format. The development of EIF was heavily related to the ideas about test scenes and the resulting EIF format is quite powerful. The created test scenes contain also visual content. A loudness calibration procedure was defined to align all scenes to a similar and comfortable/appropriate level.

To enable assessment at different test sites, a program suite called Audio Evaluation Platform (AEP) [8] was created. The video rendering is performed by Unity. The audio section is based on Max/MSP plus an API to plug in the proposed algorithms to be assessed. The specification of AEP also included the selection the hardware and procedures for the loudness calibration at each test site. For the CfP tests, subject tracking and video display used HTC Vive Pro (VR) and Microsoft HoloLens 2 (AR). Audio was output to a Focusrite Scarlett 2i2 3rd generation interface and Beyerdynamic DT-990 Pro headphone. This headphone was selected because it was available in all countries involved and the differences between individual headphones were sufficiently small. No headphone equalization was used in the CfP.

In the CfP test, only headphone rendering was tested. The final standard will encompass rendering for both headphones and loudspeakers, which was already added in the CE phase.

Finally, a proper test procedure had to be selected or developed. This will be described in the next sections.

3 Assessment of Audio Quality before MPEG-I

The target of all MPEG audio algorithms in the past had been to provide the best perceived quality at the lowest possible bitrate. Input and output to the algorithm were audio files, and

the task in a listening test was to compare input and output. Over the last 30 years, two standardized listening test method have been used.

3.1 ITU-R Recommendation BS.1116

"Methods for the subjective assessment of small impairments in audio systems" [2] is also called the "triple stimulus with hidden reference" test. The listener is presented with three stimuli: the (open) reference, and a random sequence of the reference (now called hidden reference) and the signal under test. The task of the listener is to decide which of the latter two is different from the open reference and give a score using the impairment scale (see Table 1, left side) with one decimal (scores from 1.0 up to 5.0). If the listener does not perceive a difference, it might be that the wrong stimulus is picked and so the hidden reference is downgraded. The listeners are supposed to be experts, and the correctness of the score is used to check the expertise of the listeners in a post-screening operation. Data from listeners downgrading the reference too severely and too often are discarded from further statistical analysis. This method is very sensitive to even small impairments. For comparing codecs with clearly audible distortions i.e., determining which codec provides the best (but still imperfect) quality this method is in general less reliable, because there is no explicit comparison between different codecs.

3.2 ITU-R Recommendation BS.1534

"Method for the subjective assessment of intermediate quality levels of coding systems" [3] is also called "Multiple Stimulus with hidden Reference and Anchors (MUSHRA)". The listener is presented several stimuli: the first one is the open reference, the others are a random sequence of all different signals under test, including the hidden reference and two anchors. One of the anchors is a low pass filtered version of the reference ("telephone quality": low pass at 3.5 kHz). The task of the listener is to score all stimuli using the quality scale (see Table 1, right side). One stimulus is identical to the open reference and the listeners are told to score this stimulus at 100. The listeners should be experts, and the result from the test is used to check the expertise of the listeners in a post-screening operation too. In BS.1534, there is a direct comparison between different codecs. However, the method is less sensitive to small impairments.

Impairment Scale		Quality Scale (BS.1534)	
Score	**Label**	**Score**	**Label**
1	Very annoying	0-20	Bad
2	Annoying	20-40	Poor
3	Slightly annoying	40-60	Fair
4	Perceptible, but not annoying	60-80	Good
5	Imperceptible	80-100	Excellent

Table 1: Impairment scale and quality scale. Note that on the impairment scale points are labeled, while on the quality scale intervals are labeled. In other Recommendations, the quality scale labels points.

4 Assessment of Audio Quality in MPEG-I

In MPEG-I Immersive audio, the reference only exists in the form of computer files. Listening to the content implies the usage of a renderer, but defining the best renderer is the target of the standardization process. Therefore, to use a "reference renderer" (like was done in MPEG-H 3D Audio) would have caused a bias in the listening test results. It was therefore necessary to select a listening test method which is not based on a reference. ITU-R, the organization that standardized BS.1116 and MUSHRA, provides also standards for testing audio quality without a reference:

4.1 ITU-R Recommendation BS.2132

"Method for the subjective quality assessment of audible differences of sound systems using multiple stimuli without a given reference" [4] recommends two different schemes.

4.1.1 Overall subjective quality

In Section 4.1.7, a multi-stimulus rating which is similar to ITU-R BS.1534 is specified. Neither a reference nor anchors are given. The scale used is the continuous quality scale.

4.1.2 Attribute ratings

In Section 4.1.8, a method to score different properties of a stimulus is specified. BS.2132 recommends scoring only one attribute at a time. Examples for attributes given in an informative attachment are "scene depth"," envelopment", "engulfment", "localization accuracy", "brightness", and "distortion".

4.2 ITU-R Recommendation BS.1284

"General methods for the subjective assessment of sound quality" [5] recommends many different schemes. Among them there is also a comparison of pairs using a discrete seven-grade scale.

Score	Label
3	Much better
2	Better
1	Slightly better
0	The same
-1	Slightly worse
-2	Worse
-3	Much worse

Table 2: Comparison scale

4.3 Proposed Test Methods in MPEG-I Immersive Audio

Three test schemes had been proposed: Multi-stimulus Category Rating (MuSCR), AB-Testing, and Multi Attribute Absolute Category Rating (MAACR). All three are variants of assessment methods specified in ITU-R BS.2132 and ITU-R BS.1284. To compare different proposals, these proposal systems must run in parallel in time-alignment on the AEP.

4.3.1 MuSCR

Several stimuli are presented to the test subjects with a GUI in a MUSHRA-like style. However, no open or hidden reference is given. The scoring is based on the quality scale. The obvious advantage of this method is that there is an explicit ranking between proposals. Due to the similarity to MUSHRA it was expected that the quality labels would also provide insights about the absolute quality. Pilot tests showed that the different test laboratories used the scale quite differently, disproving this assumption. With rising number of proposals, the cognitive load for test subjects is very large. The limited number of time aligned proposals which can be run by the test platform was found to be a challenge, too.

4.3.2 MAACR

Only one proposal is presented and scored. The test subjects score in four categories in parallel: basic audio quality, plausibility, externalization, and consistency. For each category only four different values (0 to 3) are possible. Proposals scoring less than 2 in at least one of the four categories for at least one scene are regarded as having insufficient quality. No final consensus has been found how the scores for the four categories, all scenes, and all test subjects can be combined to a single figure of merit.

4.3.3 AB Test

Two proposals are presented in a time-aligned way. Test subjects navigate the scene and decide which of the two sounds best. Ideally every combination of proposals must be tested by every test subject and every scene. The duration therefore raises quadratically with the number of proposals. The statistical analysis is based on the Thurston V algorithm providing a ranking of all proposals. The cognitive load for the test subjects is lower than for the other two test schemes. Only two time-aligned proposals result in a moderate computational complexity. While this method can be used to determine which proposal is the best, it does not give any hint about whether the proposal provides sufficiently good quality.

4.4 Decision for a method

Table 3 shows a comparison of pros and cons of each method.

Method	Complexity		Result
	of task	For AEP	
AB	Easy	Moderate	Ranking No absolute value
MuSCR	Well-known	High	Ranking? Absolute value?
MAACR	Unknown	Low	Rejection criteria Unclear FoM

Table 3: Summary of pros and cons of the three methods.

The task of the tests for CfP was to rank proposals. AB and MuSCR can do this, but MAACR gives no clear indication about a ranking of proposals and might even cause rejection to all proposals which was seen as a risk in CfP selection. Even after long discussions, no consensus was reached as to whether AB or MuSCR should be used for the CfP tests. Finally, a decision based on tossing a coin found consensus among all. Using an online 'random' algorithm, AB testing was selected for the CfP. Later during the CfP listening tests, it was found that for certain test scenes some combinations of proposals were too complex for the AEP. MuSCR, with more than two proposals running in parallel, would have suffered even more problems.
MuSCR is useful in the CE phase: New technologies must be statistically better for at least one scene and not worse for any other. MAACR might be the best choice for verification tests or to verify new functionalities not yet present in RM0.

4.5 Additional Considerations

4.5.1 Scene Tasks

In some pilot tests it was found that considering all requirements (like several types of occlusion, room acoustics, A/V coherence, Doppler shift, …) in one scene is too difficult for the assessors. Therefore, tests scenes have been selected where few of these effects are prominent and for each scene, instructions were written telling the assessors what to focus on ("scene task"). To cover most requirements, 14 scenes were selected.

4.5.2 Incomplete Balanced Block Design

It was foreseen that about 14 proposals have to be tested. This means 91 (14*13/2) paired comparison per listener and scene. If every listener would listen to all 14 scenes, this would have resulted in 1274 scores per listener. Pilot tests indicated that this would have resulted in several weeks of testing for each listener.

One way to reduce the amount of time per listener is to use an incomplete balanced block design [9]: The target is that every listener listens to all scenes and all proposal, but not the complete set. All pairs of proposals should occur with the same frequency at all test sites and for each listener. In general, such designs can be achieved by using Greco-Latin squares, but there are restrictions concerning the number of scenes, pairs, listeners and sites, and for most number combinations a perfect design would not lead to a reduction. Therefore, an iterative approach starting with a perfect design and controlled random selection was chosen.

4.5.3 Post-Screening of Results

Before statistical evaluation, it was necessary to check the validity of the data:

a) Assessors were allowed to reject scoring a scene. This happened rarely, mainly due to ethical reasons for a war scene.

b) The AEP monitors CPU usage. Scores where CPU usage indicated CPU overload were deleted from evaluation.

c) To check the reliability of subjective data each assessor had four self-comparison pairs. All data obtained from assessors which believed to perceive significant differences in these self-comparisons is deleted from further analysis. Data remained "in" either if all four values were in the range [-1.5:+1.5] or if three of the four values were in the range [-0.5:+0.5].

5. Statistical Analysis

Thurstone V [10] is based on absolute number of scores on pairs ("how often is proposal A scored better than proposal B"). In the AB test, the seven-point-comparison scale was used. The score figures had to be reduced to counts. The scores in the range between [-0.5:+0.5] were counted as "ties" and added as 0.5 value to both. The toolbox used for Thurstone V can only work on integer values. Therefore, all values from the score2matrix conversions were multiplied by the factor 2.

6 Results

Three tests have been carried out: Test 1 was about VR testing and encompassed the largest number of scenes and proposals. This test was also called "base line test". Test 2 addressed AR and used 7 scenes while Test 3 was about VR and 4 scenes with extended sound sources and/or multipoint HOA. In total

82 assessors participated in Test 1 at 12 test sites. The data of two listeners had to be rejected.

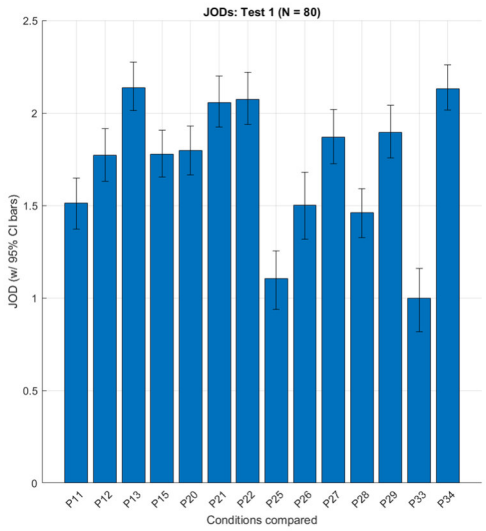

Figure 1: Just objectionable differences in test 1.

As an example of the results, the just objectionable differences (JOD) are shown in Figure 1. The winner of the base line test (Test 1) is P13. However, P21, P22, P27, P29 and P34 are statistically about the same quality. P27 came out as the winner in the category "low bitrate". All proponents had to disclose their technologies and it was found that even in the competitive phase collaboration happened and that most proposals belonged to one of two groups of proposals differing only in a few aspects. On a first glance this seems to be a waste of listening test effort, but this way it was possible to compare different alternatives in a first attempt.
Based on these results and including results from Tests 2 and 3, the technologies of P13, P27, P12 and P21 have been merged to obtain RM0 and WD0.

7 Conclusions and Lessons Learned

For comparing an extensive number of proposals, AB testing combined with incomplete balanced block design proved to be a reliable and practical solution. Bootstrapping a scale based on Thurstone V allows not only a ranking but also gives some hints about the absolute quality differences (but does not provide any hint on absolute quality).

Some data had to be rejected from the analysis because the CPU load of some pairs exceeded the capabilities of the AEP. This did not happen for all listeners and only for a few scenes: It proves that it is very difficult to predict the complexity of such renderers by an overall measurement and that this should not only based on one trajectory through the scene.

At the end it proved to be good that AB testing had been selected and not MuSCR: Based on the "CPU overruns" even with only two proposals in parallel, rendering four or even more proposals in parallel would have caused an invalidation of the test.

Acknowledgements

MPEG is a joint effort of many individuals and companies. The authors want to thank the whole audio group, and especially the convenor of MPEG Audio, Dr. Schuyler Quackenbush.

References

[1] Collection of public documents about MPEG-I Immersive Audio (2021, 2022, 2023)
https://www.mpeg.org/standards/MPEG-I/4/

[2] Recommendation ITU-R BS.1116-3 (2015) - Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems
https://www.itu.int/rec/R-REC-BS/recommendation.asp?lang=en&parent=R-REC-BS.1116

[3] Recommendation ITU-R BS.1534-3 (2015) - Method for the subjective assessment of intermediate quality level of audio systems
https://www.itu.int/rec/R-REC-BS/recommendation.asp?lang=en&parent=R-REC-BS.1534

[4] Recommendation ITU-R BS.1284-2 (2019) - General methods for the subjective assessment of sound quality
https://www.itu.int/rec/R-REC-BS/recommendation.asp?lang=en&parent=R-REC-BS.1284

[5] Recommendation ITU-R BS.2132-0 (2019) - Method for the subjective quality assessment of audible differences of sound systems using multiple stimuli without a given reference
https://www.itu.int/rec/R-REC-BS/recommendation.asp?lang=en&parent=R-REC-BS.2132

[6] MPEG-I Immersive Audio Encoder Input Format (EIF), Version 4 (2023)
https://www.mpeg.org/wp-content/uploads/mpeg_meetings/141_OnLine/w22223.zip

[7] MPEG-I Immersive Audio Augmented Reality Listener Space Description Format (LSDF) Version 2 (2023)
https://www.mpeg.org/wp-content/uploads/mpeg_meetings/141_OnLine/w22224.zip

[8] MPEG-I Immersive Audio Documentation for the Audio Evaluation Platform (AEP), Version 2 (2021)
https://www.mpeg.org/wp-content/uploads/mpeg_meetings/136_OnLine/w20920.zip

[9] Lukas Meier: Chapter 9 : Incomplete Block Designs. In ANOVA: A short Intro Using R
https://stat.ethz.ch/~meier/teaching/anova/incomplete-block-designs.html

[10] Maria Perez-Ortiz and Rafal K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments", available at
https://arxiv.org/abs/1712.03686