INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2017/N16584
January 2017, Geneva, CH

Title **MPEG-H 3D Audio Verification Test Report**
Source **Audio Subgroup**

Executive summary

MPEG-H 3D Audio is an audio coding standard developed to support coding audio as audio channels, audio objects, or Higher Order Ambisonics (HOA). MPEG-H 3D Audio can support up to 64 loudspeaker channels and 128 codec core channels, and provides solutions for loudness normalization and dynamic range control.

Four tests were conducted to assess performance of the Low Complexity Profile of MPEG-H 3D Audio. The tests covered a range of bit rates and a range of "immersive audio" use cases (i.e. from 22.2 down to 2.0 channel presentations). Seven test sites participated in the tests with a total of 288 listeners. This resulted in a data set of 15576 individual scores.

The statistical analysis of the test data resulted in the following conclusions:

- Test 1 measured performance for the "Ultra-HD Broadcast" use case, in which highly immersive audio material was coded at 768 kb/s and presented using 22.2 or 7.1+4H channel loudspeaker layouts. The test showed that at the bit rate of 768 kb/s, MPEG-H 3D Audio easily achieves "ITU-R High-Quality Emission" quality, as needed in broadcast applications.

- Test 2 measured performance for the "HD Broadcast" or "A/V Streaming" use case, in which immersive audio material was coded at three bit rates: 512 kb/s, 384 kb/s and 256 kb/s and presented using 7.1+4H or 5.1+2H channel loudspeaker layouts. The test showed that for all bit rates, MPEG-H 3D Audio achieved a quality of "Excellent" on the MUSHRA subjective quality scale.

- Test 3 measured performance for the "High Efficiency Broadcast" use case, in which audio material was coded at three bit rates, with specific bit rates depending on the number of channels in the material. Bitrates ranged from 256 kb/s (5.1+2H) to 48 kb/s (stereo). The test showed that for all bit rates, MPEG-H 3D Audio achieved a quality of "Excellent" on the MUSHRA subjective quality scale.

- Test 4 measured performance for the "Mobile" use case, in which audio material was coded at 384 kb/s, and presented via headphones. The MPEG-H 3D Audio FD binauralization engine was used to render a virtual, immersive audio sound stage for the headphone presentation. The test showed that at 384 kb/s, MPEG-H 3D Audio with binauralization achieved a quality of "Excellent" on the MUSHRA subjective quality scale.

Taken together, the tests provide evidence that the requirements set forth in the 3D Audio Call for Proposals ([1], also found in Annex 2) are fulfilled by the MPEG-H 3D Audio Low Complexity Profile.

Contents

1 Introduction

MPEG-H 3D Audio is an audio coding standard developed to support coding audio as audio channels, audio objects, or Higher Order Ambisonics (HOA). MPEG-H 3D Audio can support up to 64 loudspeaker channels and 128 codec core channels, and provides solutions for loudness normalization and dynamic range control.

Each content type (channels, objects, or HOA) can be used alone or in combination with the other ones. The use of audio channel groups, objects or HOA allows for interactivity or personalization of a program, e.g. by selecting different language tracks or adjusting the gain or position of the objects during rendering in the MPEG-H decoder.

In MPEG-H 3D Audio the format of audio program content and the coded representation that is transmitted is independent of the consumer's playback setup. The MPEG-H 3D Audio decoder renders the bitstream to a number of standard speaker configurations as well as for speakers that are not placed in the ideal positions. Binaural rendering of sound for headphone listening is also supported.

The standard may be used in a wide variety of applications including stereo and surround sound storage and transmission. Its support for interactivity and immersive sound is important to satisfy the requirements of next-generation media delivery, particularly new television broadcast systems and entertainment streaming services as well as for virtual reality content and services.

For example, in TV broadcasting, commentary or dialogue may be sent as audio objects and combined with an immersive channel bed in the MPEG-H 3D Audio decoder. This allows efficient transmission of dialogue in multiple languages and also allows the listener to adjust the balance between dialogue and other sound elements to his or her preference. This concept can be extended to other elements not normally present in a broadcast, such as audio description for the visually impaired, director's commentary, or to dialogue from participants in sporting events.

The MPEG-H 3D Audio specification is published as ISO/IEC 23008-3:2015. The requirements for the work item are shown in Annex 2. Amendment 3, specifying the Low Complexity Profile of MPEG-H 3D Audio and additional technology was published in early 2017. An integration of the base document and all amendments, as MPEG-H 3D Audio Second Edition, is expected to be published in early 2017.

Verification tests were conducted to assess the subjective quality of the Second Edition technology. Four tests were conducted to assess performance across a range of bit rates (i.e. from 768 kb/s to 48 kb/s) and a range of "immersive" use cases (i.e. from 22.2 to 2.0 channel presentations). Seven test sites participated in the tests with a total of 288 listeners. This resulted in a large data set of 15576 individual scores.

2 Listening tests

The four listening tests (Test 1, Test 2, Test 3 and Test 4) were designed to assess the performance of the Low Complexity Profile of MPEG-H 3D Audio for four important and distinct use cases in which content is broadcast to the user. A focus on broadcast delivery was chosen since the tools in the Low Complexity Profile are well matched to the broadcast scenario, although also many other applications are possible such as OTT delivery.

Test 1 assesses performance for the “Ultra HD Broadcast” use case, in which it is expected that video is Ultra HD and audio is highly immersive. Considering that such video content requires considerable bit rate, it is appropriate to allocate a proportional bit rate to audio. This test used 22.2 and 11.1 (as 7.1+4H) presentation formats, with material coded at a rate of 768 kb/s.

Test 2 assesses performance for the “HD Broadcast" or "A/V Streaming” use case, in which video has HD resolution and audio is immersive: 11.1 channel (as 7.1+4H) or 7.1 (as 5.1+2H) presentation formats. To assess codec performance for interactive content, the test contained items with multiple language tracks, that were all transmitted and the choice of the rendered language track was switched at predefined times by an automation at the decoder. For streaming and even for broadcast, there is increasing demand to deliver high-quality content at lower bitrates. In order to get a sense of the rate-distortion performance of 3D Audio, this test coded audio at three intermediate bit rates: 512 kb/s, 384 kb/s and 256 kb/s.

Test 3 assesses performance for the “High Efficiency Broadcast” use case, in which content is broadcast or streamed at very low bit rates. In order to get a sense of the rate-distortion performance of 3D Audio and to address a broader range of immersive to traditional content presentation formats, this test coded audio at three intermediate bit rates, from 256 kb/s for 5.1+2H presentation format to 48 kb/s for 2.0 presentation format.

Test 4 assesses performance for the “Mobile” use case, in which content is delivered to a mobile platform such as a smartphone. Since audio playback with such platforms is typically done via headphones, this test was conducted using headphone presentation. It used the immersive content from Test 2 (i.e. 7.1+4H and 5.1+2H presentation format) but rendered for headphone presentation using the MPEG-H 3D Audio FD binauralization engine. This permits the user to perceive a fully immersive sound stage with sound sources appropriately virtualized in the 3D space.

Listening for Test 1, Test 2 and Test 3 was conducted in acoustically isolated rooms using loudspeakers for presentation. A single subject was in the room during a given test session. Listening for Test 4 was conducted in acoustically isolated sound booths using headphones for presentation. A single subject was in the booth during a given test session.

2.1 Test methodology

BS.1116

Test 1 used the BS.1116-3 double-blind triple-stimulus with hidden reference test methodology [2]. This methodology is appropriate for assessment of systems having small impairments, and so was only used for this test in which the coding bitrate of 768 kb/s would ensure that coding artefacts would be small. The subjective response is recorded on a scale ranging from 1 to 5, with one decimal digit.

The descriptors and the score associated with each descriptor of the subjective scale are shown here:

Imperceptible	(5.0)
Perceptible, but not annoying	(4.0)
Slightly annoying	(3.0)
Annoying	(2.0)
Very annoying	(1.0)

Listener instructions for the BS.1116 test are given in Annex 6.

MUSHRA
Test 2, Test 3 and Test 4 used the MUSHRA method [3]. This methodology is appropriate for assessment of systems with intermediate quality levels. The subjective response is recorded on a scale ranging from 0 to 100, with no decimal digits.

The descriptors and the range of scores associated with each descriptor of the subjective scale are shown here:

Excellent	(80-100)
Good	(60-80)
Fair	(40-60)
Poor	(20-40)
Bad	(0-20)

Listener instructions for the MUSHRA test are given in Annex 6.

2.2 Test material

Test material was either channel-based, channel plus objects, or scene-based, as Higher Order Ambisonics (HOA) of a designated order, possibly also including objects. The number and layout of the channel-based signals is indicated as numChannels.numLFE or as numMid.numLFE + numHigh. The latter is used where there might be some confusion between a purely mid-plane layout and a mid plus high layout, e.g. 5.1+2H, where the "numHigh" is followed by "H" to indicate the high plane. The terms used in this designation are as follows:

numChannels	The total number of full-range channels, encompassing low, mid and high planes.
numLFE	The number of LFE channels
numMid	The number of mid-plane full-range channels.
numHigh	The number of high-plane full-range channels.

The filenames for each test item are given in Annex 5.

2.3 Test 1 "Ultra HD Broadcast"

The following table describes the parameters for Test 1.

Test Goal	Demonstrate ITU-R High-Quality Emission
Test Methodology	BS.1116
Presentation	Loudspeaker
Content Formats	See Test Material, Test 1 table.
Content Specialties	Switch group with 3 languages that cycles through the languages (item T1_6).
Reference	See Test Material, Test 1 table.
Test Conditions	1. Hidden Reference 2. Full decoding of all items and rendering to presentation format.
Anchor	None
Listening Position	Sweet spot

Test Items	See Test Material, Test 1 table.
Bit Rates	768 kb/s
Notes	All formats in one test Low Complexity Profile
Requirements addressed	• High Quality • Localization and Envelopment • Audio program inputs: 22.2, discrete audio objects, HOA • Interactivity

The following material was used in Test 1.

- For T1_2, item was created by rendering objects ("steps") to a 22.2 channel bed.
- For T1_5, reference was created by rendering all objects to the channel bed.
- For T1_6, reference was created by rendering the 3 commentary objects to the channel bed such that it transitions from one language to the next.
- For T1_9 and T1_11, reference was created by rendering HOA to 22.2 channels
- For T1_10 and T1_12, reference was created by rendering HOA to 7.1+4 channels.

Item	**Content Format**	**Presentation Format**	**Item Name**	**Item Description**
T1_1	22.2	22.2	Funk	Drums, guitar, bass
T1_2	22.2	22.2	Rain with steps	Rain with steps (steps as obj)
T1_3	22.2	22.2	Swan Lake	Tchaikovsky with full orchestra
T1_4	22.2	22.2	This is SHV	Trailer for 8K Super Hi-Vision
T1_5	7.1+4H + 3 obj	7.1+4H	Sintel Dragon Cave (3 obj)	Fighting film scene with score
T1_6	7.1+4H + 3 obj	7.1+4H	DTM Car Race (3 obj, commentary languages)	Car race with 3 commentaries in 3 different languages
T1_7	7.1+4H	7.1+4H	Birds Paradise	Ambience with birds
T1_8	7.1+4H	7.1+4H	Musica Floria	String ensemble recorded in medieval church
T1_9	HOA + 2 obj	22.2	FTV Yes (2 obj, English language)	Movie scene with 2 languages
T1_10	HOA + 1 obj + 1 LFE	7.1+4H	DroneObj (1 obj, 1 LFE)	Drama with object
T1_11	HOA	22.2	Moonshine	A capella ensemble
T1_12	HOA	7.1+4H	H_12_Radio	Guitars

2.4 Test 2 "HD Broadcast" or "A/V Streaming"

The following table describes the parameters for Test 2.

Test Goal	Demonstrate MUSHRA "Excellent" (80+)
Test Methodology	MUSHRA

Presentation	Loudspeaker
Content Formats	See Test Material, Test 2 table.
Content Specialties	Switch group with 2 languages that cycles through the languages (item T2_6).
Reference	See Test Material, Test 2 table.
Test Conditions	1. Hidden Reference 2. 3D Audio at 512 kb/s 3. 3D Audio at 384 kb/s 4. 3D Audio at 256 kb/s 5. Anchor 1 6. Anchor 2
Anchor	Anchor 1: original, LP filtered, 7.0 kHz Anchor 2: original, LP filtered, 3.5 kHz
Listening Position	Sweet spot
Test Items	See Test Material, Test 2 table.
Bit Rates	Three bit rates as shown above
Notes	All formats in one test Low Complexity Profile
Requirements addressed	• High Quality • Localization and Envelopment • Audio program inputs: channel-based PCM, discrete audio objects, HOA • Interactivity

The following material was used in Test 2.

- For T2_1, item was created by rendering objects to a 7.1+4H channel bed.
- For T2_2, item was created by rendering the 3 commentary objects to the channel bed such that it transitions from one language to the next.
- For T2_5, reference was created by rendering object to 5.1+2H channel bed.
- For T2_6, reference was created by rendering the 2 commentary objects to the channel bed such that it transitions from the English commentary to the German commentary.
- For HOA items, reference was created by rendering to 7.1+4H channels.

Item	**Content Format**	**Presentation Format**	**Item Name**	**Item Description**
T2_1	7.1+4H	7.1+4H	Sintel Dragon Cave	Fighting film scene with score
T2_2	7.1+4H	7.1+4H	DTM Car Race	Car race with 3 commentaries in 3 different languages
T2_3	7.1+4H	7.1+4H	Birds Paradise	Ambience with birds
T2_4	7.1+4H	7.1+4H	Musica Floria	String ensemble recorded in medieval church
T2_5	5.1+2H + 3 obj	5.1+2H	Sintel Dragon Cave	Fighting film scene with score
T2_6	5.1+2H + 2 obj	5.1+2H	Handball Commentary	Sports with commentaries in 2

				different languages
T2_7	5.1+2H	5.1+2H	Blug Hendrix Beat	Live rock concert
T2_8	5.1+2H	5.1+2H	Song World Percussion	Pop Music with drums
T2_9	HOA	7.1+4H	Moonshine	A capella
T2_10	HOA	7.1+4H	H_12_Radio	Guitars
T2_11	HOA	7.1+4H	Drone	Drama
T2_12	HOA	7.1+4H	H_07_Vocal1	Female voice with piano and orchestra

2.5 Test 3 "High Efficiency Broadcast"

The following table describes the parameters for Test 3.

Test Goal	Demonstrate MUSHRA "Good" quality at low bit rates
Test Methodology	MUSHRA
Presentation	Loudspeaker
Content Formats	See Test Material, Test 3 table.
Content Specialties	None
Reference	See section on Test 3 Material, above.
Test Conditions	1 Hidden Reference 5.1+2H 5.1 2.0 HOA 2 3D Audio 256 kb/s 180 kb/s 80 kb/s 256 kb/s 3 3D Audio 192 kb/s 144 kb/s 64 kb/s 192 kb/s 4 3D Audio 144 kb/s 128 kb/s 48 kb/s 144 kb/s 5 Anchor 1 6 Anchor 2
Anchor	Anchor 1: original, LP filtered, 7.0 kHz Anchor 2: original, LP filtered, 3.5 kHz
Listening Position	Sweet spot
Test Items	See Test Material, Test 3 table.
Bit Rates	As in Test Conditions row of this table.
Notes	All formats in one test Low Complexity Profile No interactivity No dynamic objects
Requirements addressed	• High Quality • Localization and Envelopment • Audio program inputs: channel-based PCM, discrete audio objects, HOA

The following material was used in Test 3.

- For T3_1 and T3_2, item was created by rendering all objects to a 5.1+2H channel bed.
- For T3_2, only English commentary was used.
- For all HOA items, reference was created by rendering to 5.1+2H channels.
- For T3_10 and T3_12, item was created by truncating HOA originals to third order HOA prior to rendering.

- T3_11 was used as is, i.e. HOA 6th order.

Item	Content Format	Presentation Format	Item Name	Item Description
T3_1	5.1+2H	5.1+2H	Sintel Dragon Cave	Fighting film scene with score
T3_2	5.1+2H	5.1+2H	Handball Commentary	Sports with commentary
T3_3	5.1+2H	5.1+2H	Blug Hendrix Beat	Live rock concert
T3_4	5.1	5.1	Mancini	Movie score with brass
T3_5	5.1	5.1	Bach 565	Bach Toccata d minor
T3_6	5.1	5.1	Sedambonjou Salsa	Latin music with brass and percussions
T3_7	2.0	2.0	Susanne Vega (te8)	Suzanne Vega, Tom's Diner
T3_8	2.0	2.0	Tracy Chapman (te9)	Tracy Chapman
T3_9	2.0	2.0	Hockey	Hockey Game
T3_10	HOA	5.1+2H	Moonshine	A capella
T3_11	HOA	5.1+2H	Drone	Drama
T3_12	HOA	5.1+2H	H_07_Vocal1	Female voice with piano and orchestra

Note: Items T3_5, T3_6 and T3_9 were kindly provided by EBU.

2.6 Test 4 "Mobile"

The following table describes the parameters for Test 4.

Test Goal	Demonstrate MUSHRA "excellent" (80+)
Test Methodology	MUSHRA
Presentation	Headphones
Content Formats	Same as in Test 2, "HD Broadcast" or "A/V Streaming"
Content Specialties	None
Reference	Channels: PCM original item processed by BRIR as full convolution. HOA: Reference rendering of the HOA to the Presentation Format, then processed by BRIR as full convolution. Objects: If items contain objects, the objects are rendered to Presentation Format and then processed by BRIR as full convolution. BRIR are the same BRIR as was used in MPEG-H 3D Audio CfP
Test Conditions	1. Hidden Reference 2. C/O: MPEG-H using FD binauralization engine HOA: MPEG-H using FD binauralization engine 3. Anchor 1 4. Anchor 2
Anchor	Anchor 1: Anchor 1 from Test 2, then processed by BRIR Anchor 2: Anchor 2 from Test 2, then processed by BRIR
Listening Position	N/A
Test Items / Bit	Use 384 kb/s bitstreams from Test 2

Rates	
Restrictions	None
Notes	All formats in one test
Requirements addressed	• High Quality • Localization and Envelopment • Audio program inputs: channel-based PCM, discrete audio objects, HOA • Rendering for Headphone Listening • HRTF Personalization

Test 4 used the same material as Test 2. More specifically, in Test 4 the 3D Audio decoder processed the Test 2 bitstreams to create a binauralized stereo result. The binauralization used a Binaural Room Impulse Response (BRIR), specifically, the same BRIR as was used in the MPEG-H 3D Audio Call for Proposals [1]. This BRIR was recorded in the Mozart listening room at Fraunhofer IIS.

3 Test plan

3.1 Preparation of original and processed items

Original items were provided by ARL, EBU, ETRI, Fraunhofer IIS, FTV, NHK, Orange and Qualcomm. They were limited to not more than 20 seconds duration and were edited to have "fade-in" and "fade-out" at beginning and end.

All channel and channel plus object test items were processed, i.e. encoded/decoded and low-pass filtered, by Fraunhofer IIS. All HOA and HOA plus object test items were processed, i.e. encoded/decoded and low-pass filtered, by Qualcomm.

3.2 Listening labs

The following table shows the listening labs that participated in each listening test. The number of subjects participating from each lab in a given test is shown in the table entries; a blank entry indicates no participation. The total number of listeners in each test is shown in the last line of the table.

Test	**Test 1**	**Test 2**	**Test 3**	**Test 4**
ETRI		12		12
FhG-IIS	24	24	29	28
NHK	18	18	18	
Nokia			10	12
Orange				9
Qualcomm	16	15	16	16
Sony	11			
Total	**69**	**69**	**73**	**77**

For Test 1, Test 2 and Test 3, the listening labs all had high-quality listening rooms that were calibrated to conform to the criteria set forth in BS.1116 and also calibrated to be perceptually similar to each other. Hence, the test lab subjective results can be pooled together for each of the tests. The loudspeaker positions used when presenting the various test

item is shown in Table 2 of Annex 5, specifically the loudspeaker azimuth (A+000) and elevation (E+00) angles are shown under the heading "Label."

For Test 4, the listening labs used acoustically isolating sound booths and high-quality headphones.

4 Statistical Analysis and Test Results

4.1 Listener post-screening

Test 1

Test 1 used the BS.1116 test methodology [2]. For each listener in Test 1, post-screening of listener responses was based on the listener's ability to correctly differentiate the Hidden Reference from the System under Test, which is the procedure recommended in BS.1116-3. The exact procedure used is described in Annex 3.

The post-screening procedures computes the statistic T_i which is the 95% point of the cumulative distribution of the listener Diff Grades, which are assumed to have the Student t distribution. If $T_i > 0$ for the listener i, then we conclude, with a 95% level of significance, that the listener cannot reliably differentiate between the Hidden Reference and the System under Test, and listener responses for the 12 test items are removed from consideration.

Test 2, Test 3, Test 4

Test 2, Test 3 and Test 4 used the MUSHRA test methodology [3]. For each listener in each test, post-screening of listener responses was based on scores for Hidden Reference and Low Pass filtered anchors. The procedure is as follows:

If, for any test item in a given test, either of the following criterion are *not* satisfied:

- The listener score for the Hidden Reference is greater than or equal to 90 (i.e. HR >= 90)
- The listener score for the Hidden Reference, the 7.0 kHz lowpass anchor and the 3.5 kHz lowpass anchor are monotonically decreasing (i.e. HR >= LP70 >= LP35).

then all listener responses in that test are removed from consideration.

Post-Screening Result

After applying these listener post-screening rules, the number of listeners remaining for each test is shown in the following table.

Test	Test 1	Test 2	Test 3	Test 4
After Post-Screening	35	43	44	68

After applying post-screening there were at least 35 listeners for every test. This number far exceeds the BS.1116-3 and BS.1534-3 recommendations of at least 20 listeners per test.

4.2 Overview

Statistical analysis was performed on subjective scores remaining after listener post-screening. Details of the statistical analysis are given in Annex 3. For Test 1, a Diff Grade was computed (as Hidden Reference – System under Test scores) and statistics were computed on the Diff Grade. In addition, statistical analysis was performed on absolute

scores for Hidden Reference and the System under Test. For Test 2, Test 3 and Test 4, statistics were computed on the absolute MUSHRA scores.

The tables in this section show, for each System under Test (Sys), the mean score (Mean) as averaged over all listeners (after post-screening) and all test items. For each result, the 95% confidence interval on the mean score was computed, and the table shows the upper (High) and lower (Low) limits of the 95% confidence interval.

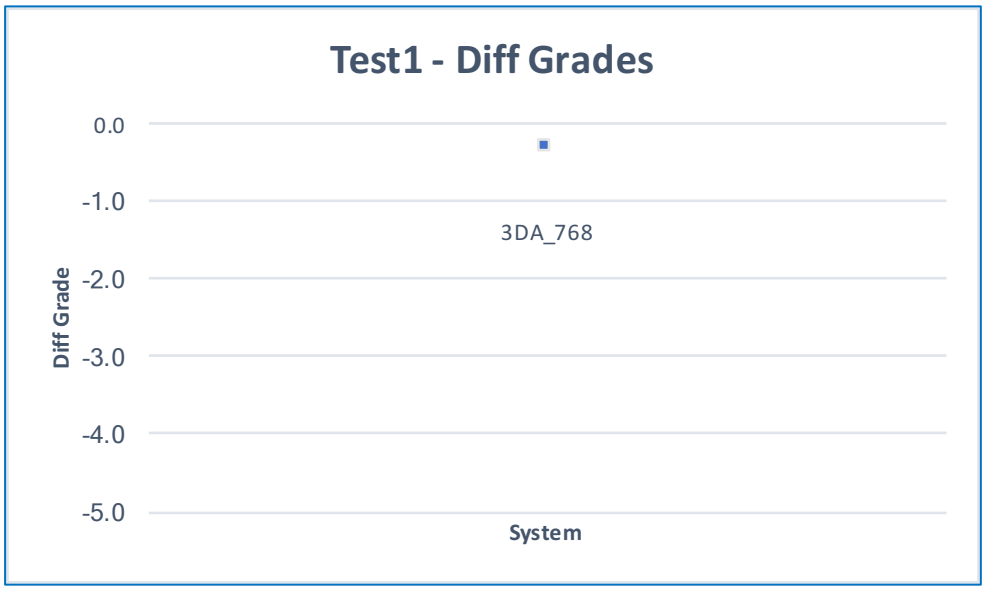
Note that the 95% confidence interval is shown in every plot, but when retaining the full subjective scale, the interval is obscured by the mark used to indicate the mean value. However, 95% confidence intervals are shown in the tabular presentation of scores.

4.3 Test 1 "Ultra HD Broadcast"

The following table shows the mean score for 3D Audio system operating at 768 kb/s (3DA_768) and the associated high and low 95% confidence interval limits on the mean.

Sys	High	Low	Mean
3DA_768	-0.27	-0.35	-0.31

The following is a plot of the mean score and 95% confidence interval. The confidence interval is plotted, but is so small that it is within the size of the marker used for the mean.



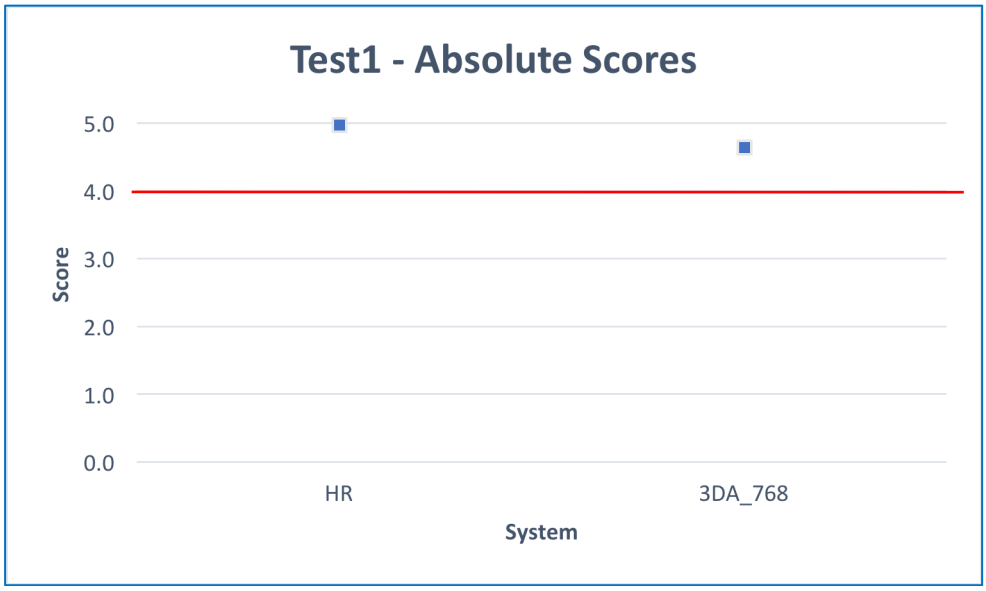
The following table and plot show the mean score for 3D Audio system operating at 768 kb/s (3DA_768), the Hidden Reference (HR) and the associated high and low 95% confidence interval limits on the mean for each condition.

For the 3DA_768, the absolute score is not lower that 4.6 at the 95% level of confidence, which is well above the 4.0 limit recommended in ITU-R BS.1548-4 for "High-quality emission" for broadcast applications (indicated by red line in the plot). Recommendation ITU-R BS.1548-4, Section 2.1.1.1 "High-quality emission" states "Ideally, the quality of the sound reproduced after decoding will be subjectively similar to the original signal for most types of audio programme material. Using the triple stimuli double blind with hidden

reference test, described in Recommendation ITU-R BS.1116, this requires mean values consistently higher than 4 on the Recommendation ITU-R BS.1116 5-grade impairment scale at the reference listening position."

Sys	High	Low	Mean
HR	5.05	4.90	4.98
3DA_768	4.67	4.61	4.64

The following is a plot of the mean scores and 95% confidence intervals. The confidence intervals are plotted, but are so small that they are within the size of the marker used for the mean. The red line shows the ITU-R requirement for "high-quality emission," i.e. mean value of 4.0.

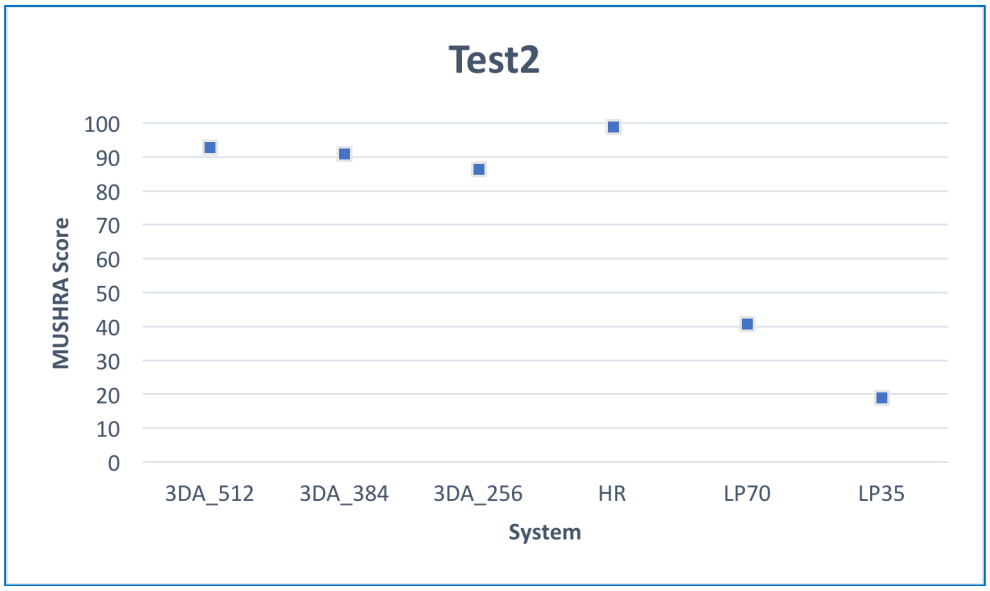


4.4 Test 2 "HD Broadcast" or "A/V Streaming"

The following table shows the mean score for 3D Audio operating at 512 kb/s (3DA_512), 384 kb/s (3DA_384), 256 kb/s (3DA_256), the Hidden Reference (HR), the 7.0 kHz low pass anchor (LP70) and 3.5 kHz low pass anchor (LP35), and the associated high and low 95% confidence interval limits on the mean for each condition.

Sys	High	Low	Mean
3DA_512	93.44	92.15	92.79
3DA_384	91.60	90.27	90.93
3DA_256	87.39	85.54	86.47
HR	99.12	98.70	98.91
LP70	41.88	39.93	40.91
LP35	19.65	18.32	18.99

The following is a plot of the mean scores and 95% confidence intervals. The confidence intervals are plotted, but are so small that they are within the size of the marker used for the mean.

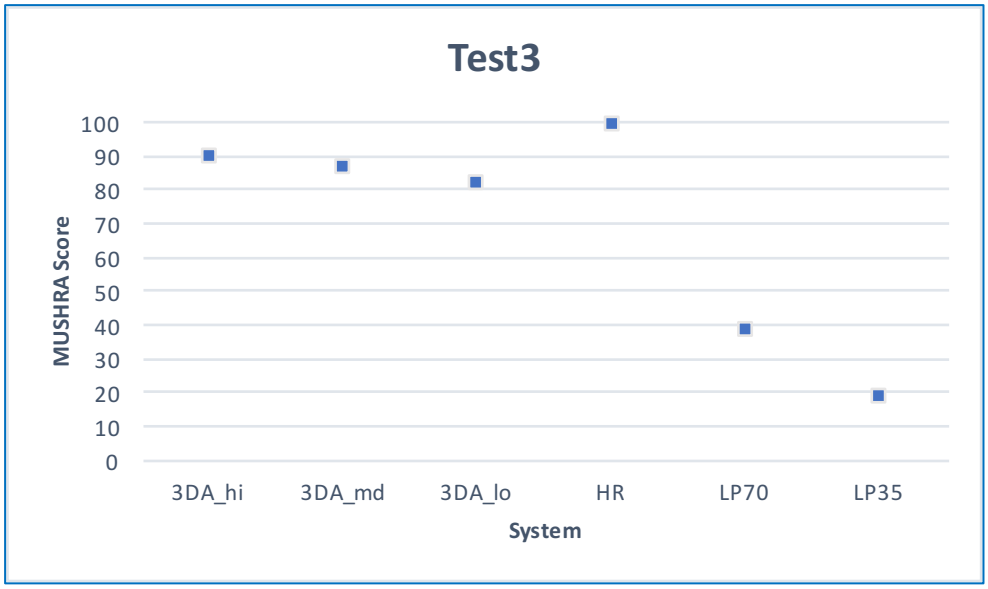


4.5 Test 3 "High Efficiency Broadcast"

The following table shows the mean score for 3D Audio operating at three bit rates: 3DA_hi, 3DA_mid, 3DA_lo, the Hidden Reference (HR), the 7.0 kHz low pass anchor (LP70) and 3.5 kHz low pass anchor (LP35), and the associated high and low 95% confidence interval limits on the mean for each condition. The specific bit rates for each test item for each of the three rates (hi, mid, lo) are given in the table in Section 2.5.

Sys	High	Low	Mean
3DA_hi	91.00	89.08	90.04
3DA_md	87.63	85.35	86.49
3DA_lo	83.46	80.80	82.13
HR	99.36	98.99	99.18
LP70	39.47	36.87	38.17
LP35	19.71	17.84	18.77

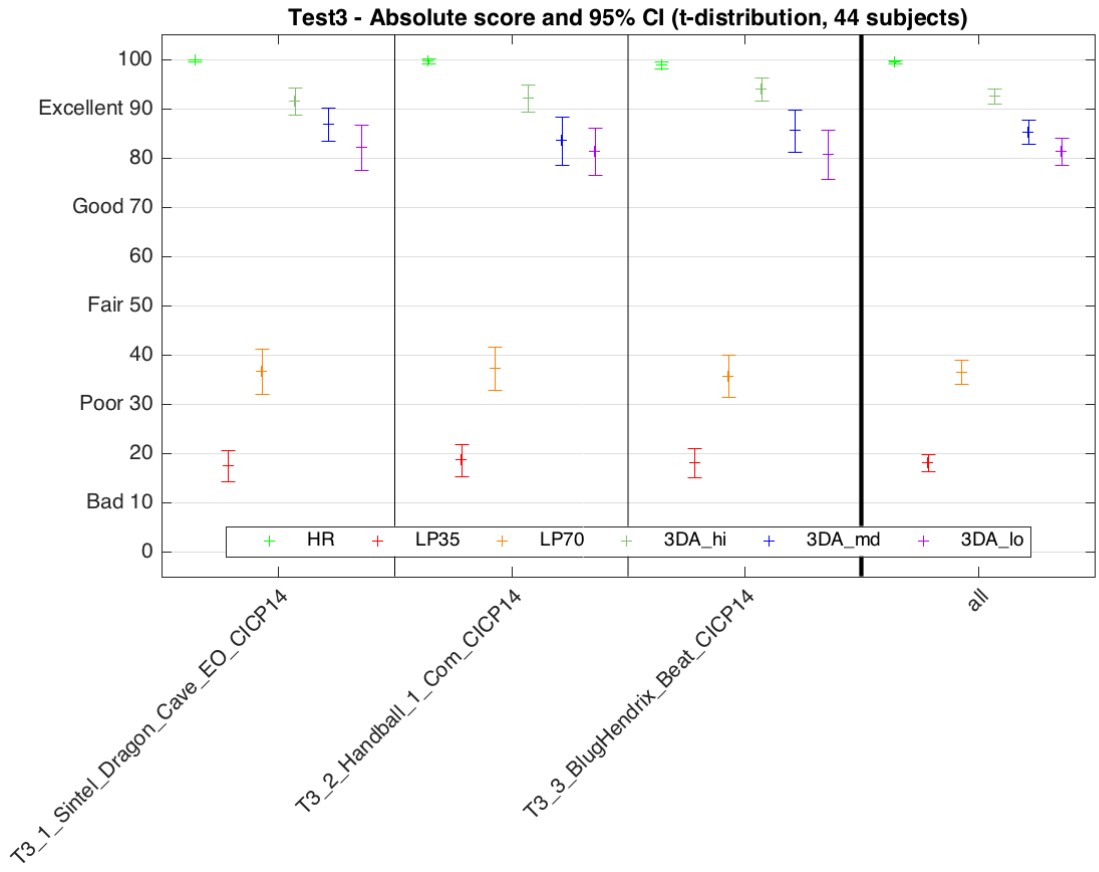
The following is a plot of the mean scores and 95% confidence intervals. The confidence intervals are plotted, but are so small that they are within the size of the marker used for the mean.



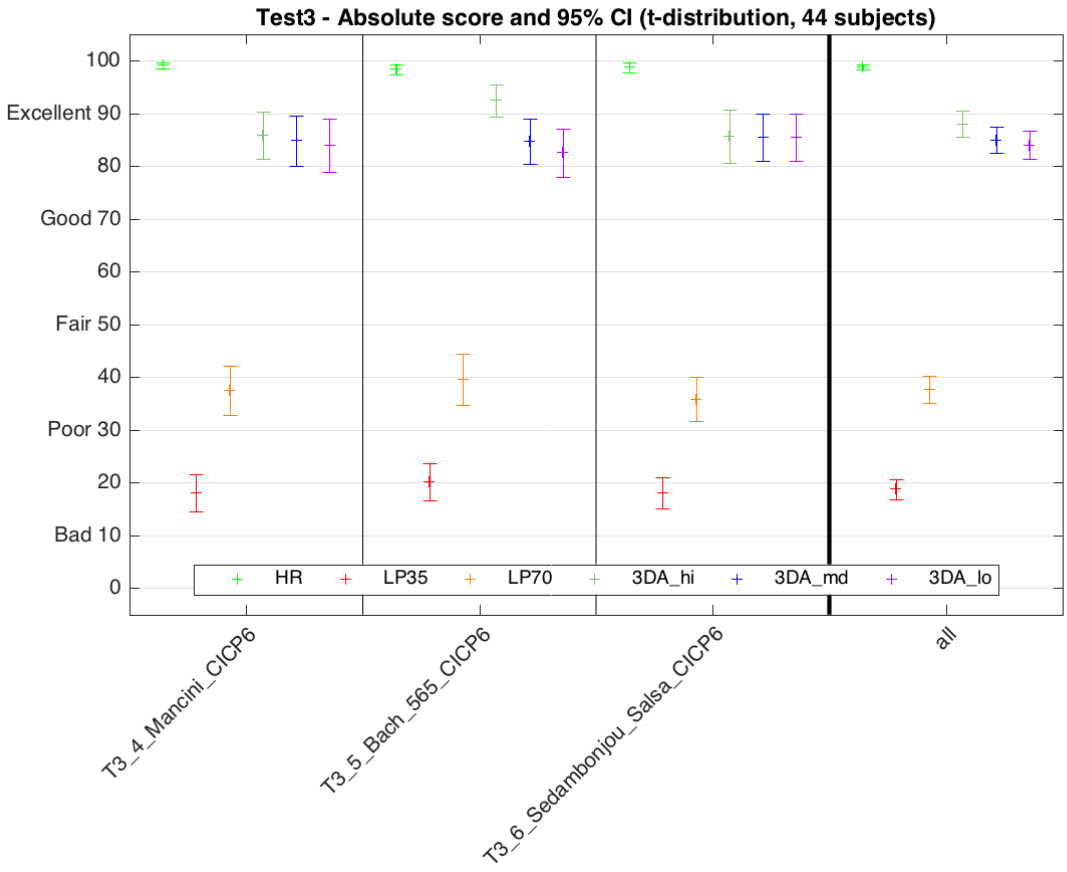
Since this test used a range of content formats for the test items and coded each format with a range of bit rates, the following table and plots present the performance of 3D Audio for each content format for the three (hi, mid, lo) coding bit rates.

Content	High Rate	Mid Rate	Low Rate
Stereo	90.60 ± 1.68	88.68 ± 1.98	81.83 ± 2.81
5.1	88.00 ± 2.47	85.02 ± 2.52	84.02 ± 2.63
5.1+2	92.50 ± 1.50	85.23 ± 2.36	81.29 ± 2.71
HOA @ 5.1+2	89.05 ± 1.87	87.02 ± 2.26	81.39 ± 2.56

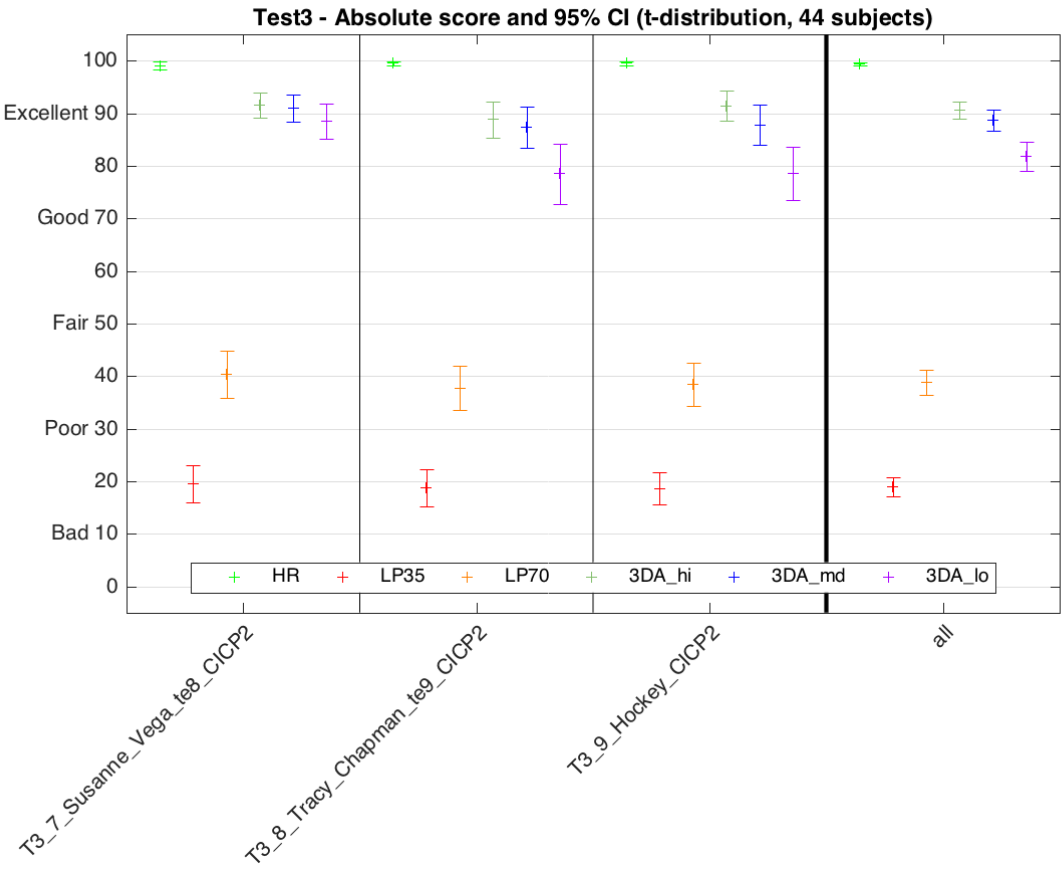
The following plot shows the performance for 5.1+2H layout (CICP 14) immersive content.



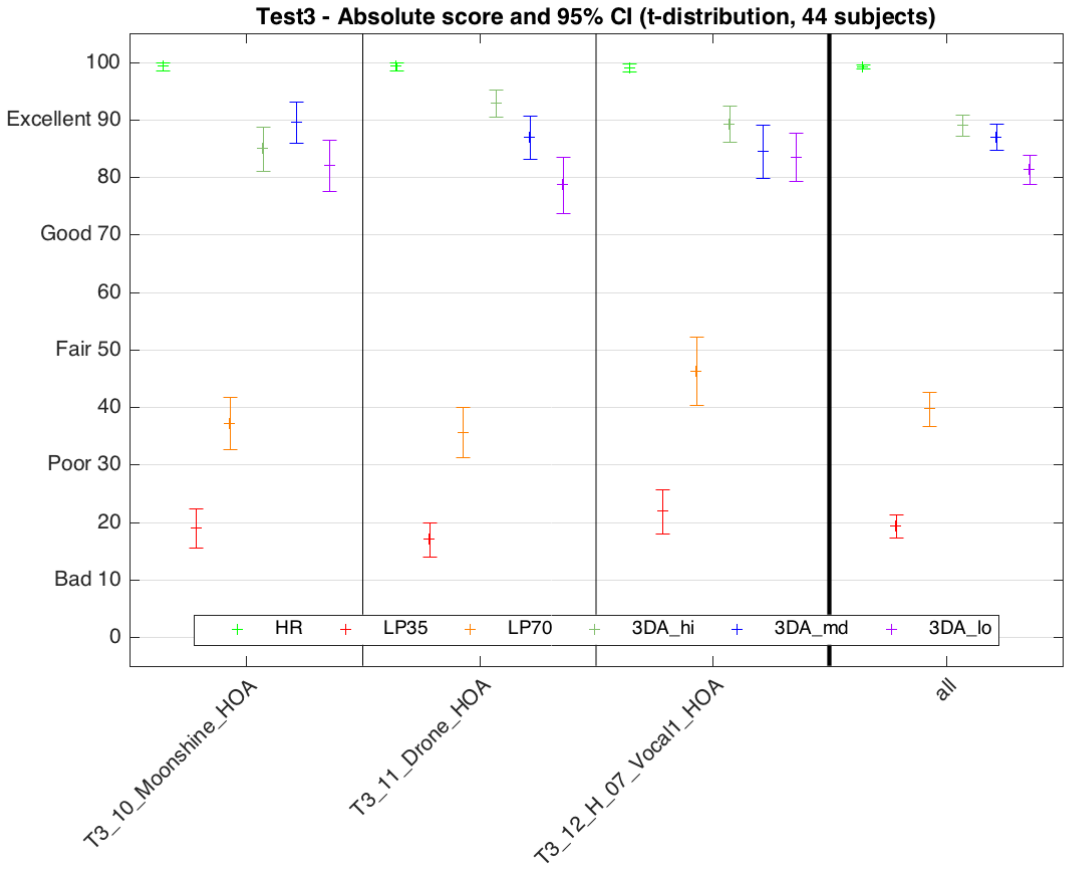
The following plot shows the performance for 5.1 layout (CICP 6) content.



The following plot shows the performance for stereo (CICP 2) content.



The following plot shows the performance for HOA content.

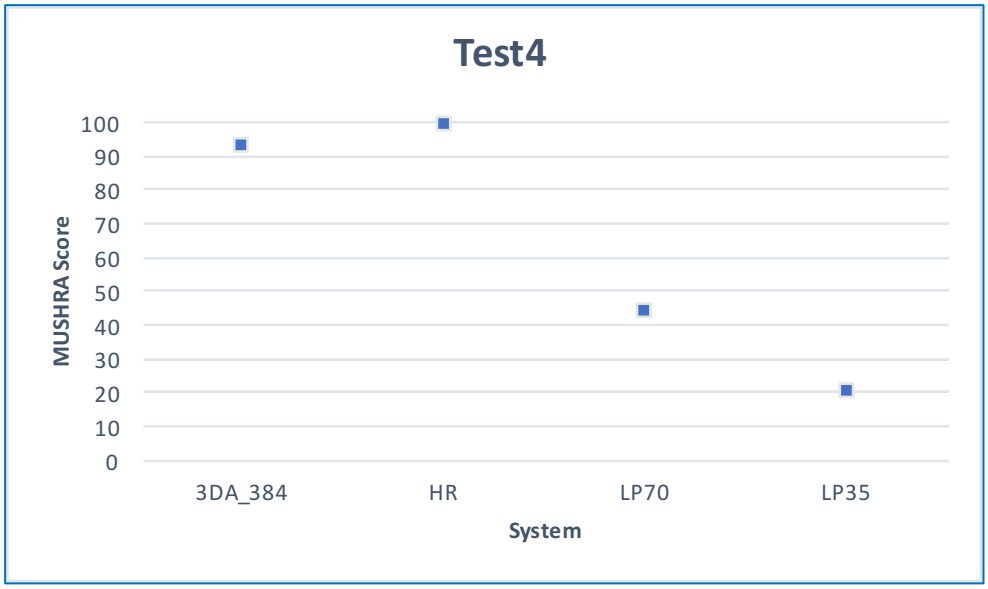


4.6 Test 4 "Mobile"

The following table shows the mean score for 3D Audio operating at 384 kb/s (3DA_384), the Hidden Reference (HR), the 7.0 kHz low pass anchor (LP70) and 3.5 kHz low pass anchor (LP35), and the associated high and low 95% confidence interval limits on the mean for each condition.

Sys	High	Low	Mean
3DA_384	93.76	92.76	93.26
HR	99.37	99.09	99.23
LP70	44.71	42.76	43.73
LP35	20.95	19.50	20.22

The following is a plot of the mean scores and 95% confidence intervals. The confidence intervals are plotted, but are so small that they are within the size of the marker used for the mean.



5 Conclusion

This report provides details on four tests that were conducted to assess the performance of the Low Complexity Profile of MPEG-H 3D Audio. The tests covered a range of bit rates and a range of "immersive audio" use cases (i.e. from 22.2 down to 2.0 channel presentations).

The statistical analysis of the test data resulted in the following conclusions:

- Test 1 measured performance for the "Ultra-HD Broadcast" use case, in which highly immersive audio material was coded at 768 kb/s and presented using 22.2 or 7.1+4H channel loudspeaker layouts. The test showed that at the bit rate of 768 kb/s, MPEG-H 3D Audio easily achieves "ITU-R High-Quality Emission" quality, as needed in broadcast applications.

- Test 2 measured performance for the "HD Broadcast" or "A/V Streaming" use case, in which immersive audio material was coded at three bit rates: 512 kb/s, 384 kb/s and 256 kb/s and presented using 7.1+4H or 5.1+2H channel loudspeaker layouts. The test showed that for all bit rates, MPEG-H 3D Audio achieved a quality of "Excellent" on the MUSHRA subjective quality scale.

- Test 3 measured performance for the "High Efficiency Broadcast" use case, in which audio material was coded at three bit rates, with specific bit rates depending on the number of channels in the material. Bitrates ranged from 256 kb/s (5.1+2H) to 48 kb/s (stereo). The test showed that for all bit rates, MPEG-H 3D Audio achieved a quality of "Excellent" on the MUSHRA subjective quality scale.

- Test 4 measured performance for the "Mobile" use case, in which audio material was coded at 384 kb/s, and presented via headphones. The MPEG-H 3D Audio FD binauralization engine was used to render a virtual, immersive audio sound stage for the headphone presentation. The test showed that at 384 kb/s, MPEG-H 3D Audio with binauralization achieved a quality of "Excellent" on the MUSHRA subjective quality scale.

Taken together, the tests provide evidence that the requirements set forth in the 3D Audio Call for Proposals ([1], also found in Annex 2) are fulfilled by the MPEG-H 3D Audio Low Complexity Profile.

6 References

[1] N13411, "Call for Proposals for 3D Audio." Available at http://mpeg.chiariglione.org/standards/mpeg-h/3d-audio

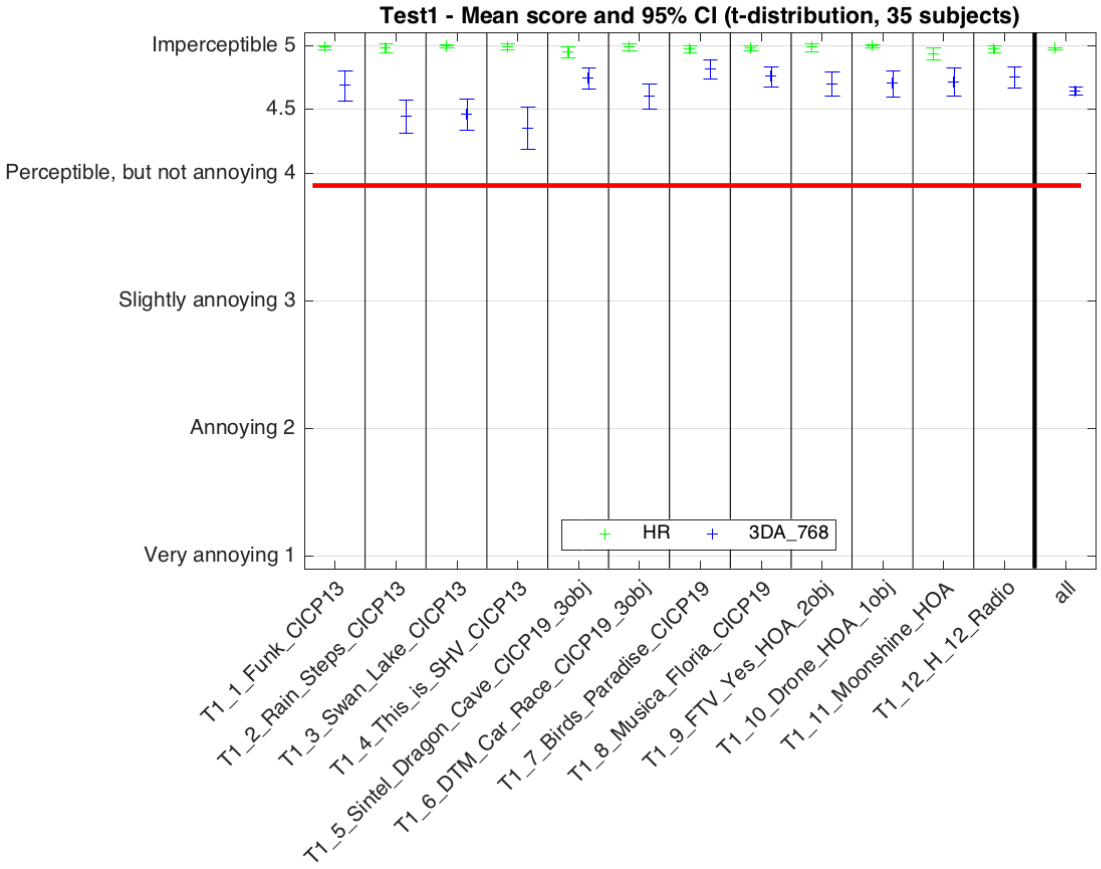
[2] ITU-R Recommendation BS.1116-3 (02/2015), "Methods for the subjective assessment of small impairments in audio systems."

[3] ITU-R Recommendation BS.1534-3 (10/2015), "Method for the subjective assessment of intermediate quality level of coding systems," also known as "MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)."

Annex 1 Performance for individual test items

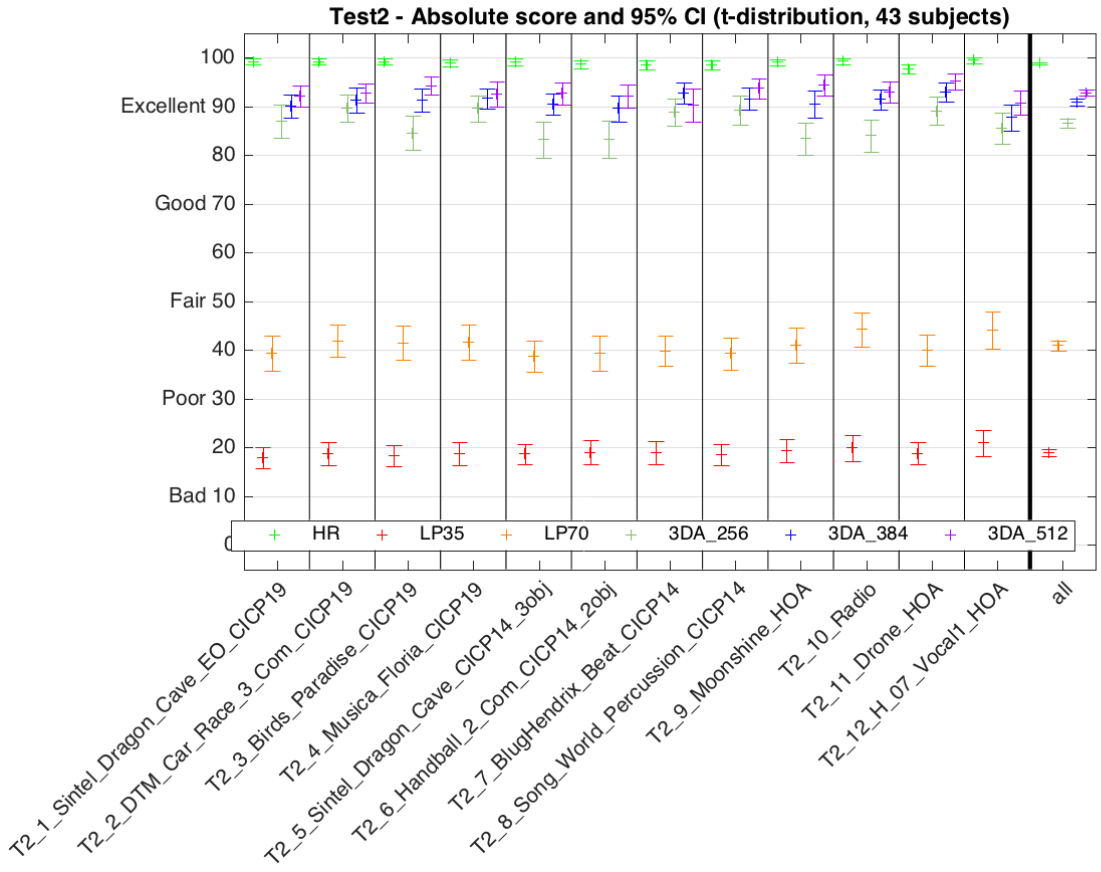
Test 1

This test used the BS.1116 test methodology. Test items were coded at 768 kb/s and test material was played out as 22.2 and 7.1+4H channel presentations. For all test items, the absolute score is above 4.0 at the 95% level of confidence, which meets the ITU-R BS.1548-4 recommendation for "High-quality emission" for broadcast applications



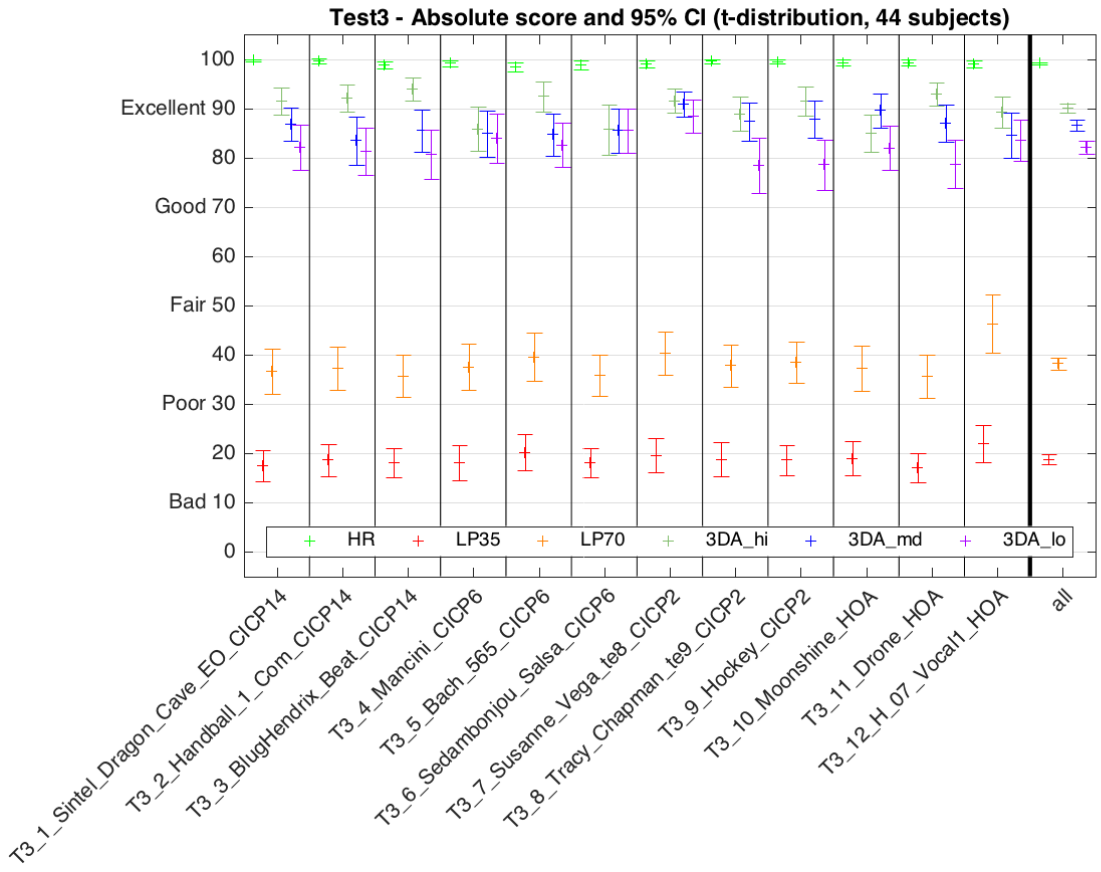
Test 2

This test used the MUSHRA test methodology. Test items were coded at 512, 384 and 256 kb/s and test material played out as 11.1 and 5.1+2H channel presentations.



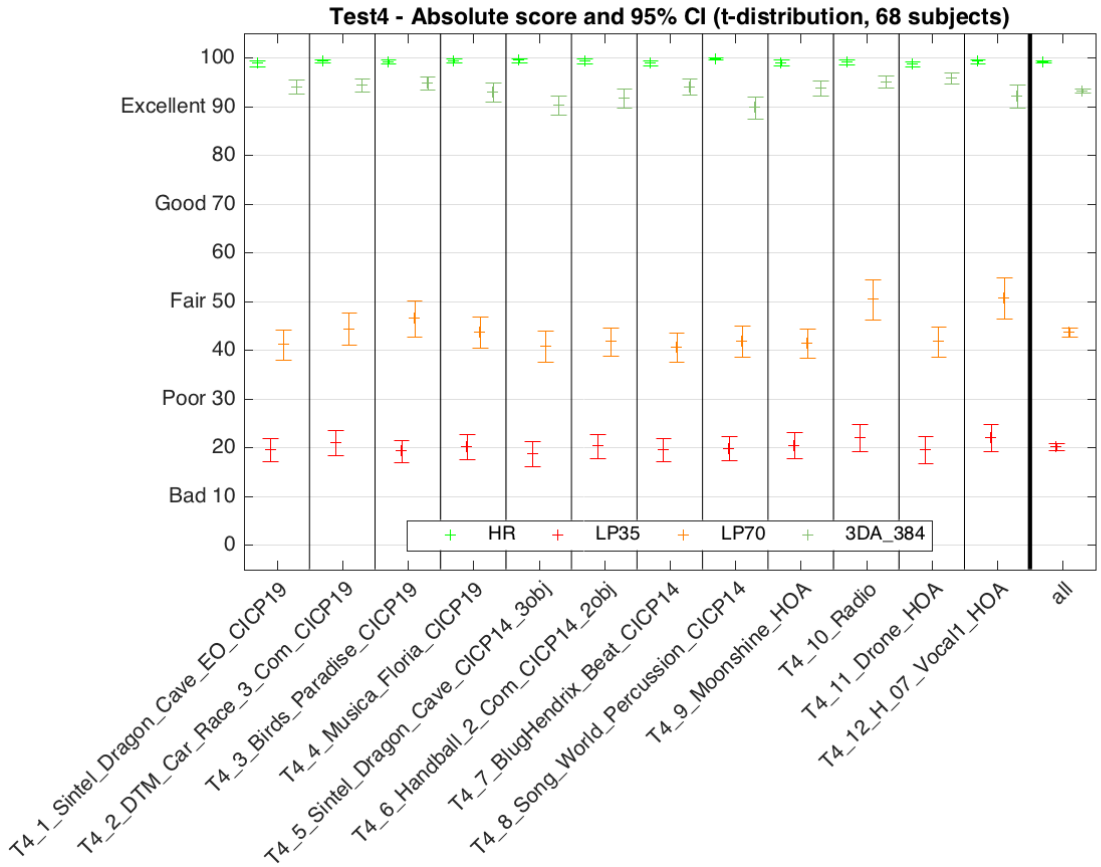
Test 3

This test used MUSHRA test methodology. Test items were coded at various rates, from 256 kb/s for 5.1+2H channel material to 48 kb/s for 2.0 channel material. See Section 2.5 for complete information.



Test 4

This test used MUSHRA test methodology. The test used the 384 kb/s material from Test 2, but used the 3D Audio FD binauralization to virtualize for presentation via headphones.


Test4 - Absolute score and 95% CI (t-distribution, 68 subjects)
100
Excellent 90
80
Good 70
60
Fair 50
40
Poor 30
20
Bad 10
0
HR
LP35
LP70
3DA_384
T4_1_Sintel_Dragon_Cave_EO_CICP19
T4_2_DTM_Car_Race_3_Com_CICP19
T4_3_Birds_Paradise_CICP19
T4_4_Musica_Floria_CICP19
T4_5_Sintel_Dragon_Cave_CICP14_3obj
T4_6_Handball_2_Com_CICP14_2obj
T4_7_BlugHendrix_Beat_CICP14
T4_8_Song_World_Percussion_CICP14
T4_9_Moonshine_HOA
T4_10_Radio
T4_11_Drone_HOA
T4_12_H_07_Vocal1_HOA
all

Annex 2 Requirements for MPEG-H 3D Audio work item

The MPEG-H 3D Audio standard shall fulfill all Primary Requirements. Favorable consideration will be given to technology that additionally fulfills Secondary Requirements.

Primary Requirements:

- **High quality**: For high-quality applications, the quality of decoded sound shall scale up to be perceptually transparent with increasing bit rate.
- **Localization and Envelopment**: Accurate sound localization shall be supported and the sense of sound envelopment shall be very high within a targeted listening area. Perceived audio sound source distance shall be supported as a part of sound localization.
- **Rendering on setups with fewer loudspeakers**: the bitstream/compressed representation shall support decoding/rendering with a lower number of loudspeakers than are present in the loudspeaker setup used for the reference rendering of the program material. The decoded/rendered output signal shall have highest possible subjective quality relative to the reference rendering.
- **Flexible Loudspeaker Placement**: the bitstream/compressed representation shall be able to be decoded and rendered to a setup in which loudspeakers are in alternate (i.e. non-standard) positions and possibly fewer positions while providing highest possible subjective quality.
- **Latency:** technology shall have sufficiently low latency to be able to support live broadcasts (e.g. live sporting events). One-way algorithmic latency shall not exceed 1 second.
- **Audio program inputs to envisioned 3D Audio standard:**
    - Shall accept channel-based PCM signals of at least 22 full-bandwidth channels and 2 LFE channels (i.e. 22.2) that are configured to directly feed reproduction loudspeakers.
    - May accept discrete audio objects as PCM signals with associated rendering/position/scene information.
    - May accept PCM signals that use Higher Order Ambisonics representation.
- **Rendering for Headphone Listening**
    - The standard shall be able to do binaural rendering for headphones.
    - **HRTF Personalization:** Decoder shall support a normative format for reading in a user-specified Head-Related Transfer Function (HRTF) for spatialization, e.g. for headphone listening.

Secondary Requirements

- **Computational complexity** should be appropriate for the target application scenario. For example, for broadcasting it is appropriate that decoder/rendering have low computational complexity, while encoder complexity is not critical.
- **Interactivity**: Interactive modification of the sound scene rendered from the coded representation, e.g. by control of audio objects prior to rendering, may be supported for use in personal interactive applications.

Annex 3 Post-screening and statistical analysis

A.1 Post-screening analysis

A post-screening procedure was applied to listener data in all tests to assess the subjects' reliability.

BS.1116
Test 1 used the BS.1116 test methodology. For each listener in the test, post-screening was based on the listener's ability to correctly differentiate between the Hidden Reference and the System under Test, which is the procedure recommended in BS.1116-3.

The first step is to calculate Diff Grades (d) for each listener trial

$$d_{i,j} = SuT_{i,j} - HR_{i,j}$$

where

$d_{i,j}$ is Diff Grade
$SuT_{i,j}$ is the score for the System under Test
$HR_{i,j}$ is the score for the Hidden Reference

for

subject *i* and test item *j*.

Note that if the listener ability to correctly differentiate between the Hidden Reference and the System under Test, the listener's Diff Grades are typically less than zero since the listener should score the Hidden Reference to 5.0 and the System under Test to less than 5.0.

A single-sided test, in which the Diff Grade has the Student t distribution, is used to assess the ability of a given listener to correctly differentiate between Hidden Reference and the System under Test. We compute the statistic T_i:

$$T_i = \overline{d_i} + \left(t_{\propto,n-1}\right)\left(S_i/\sqrt{n}\right)$$

where

$t_{\propto,n-1}$ is the inverse Student t distribution value, that is the point in the Student t distribution for which α probability is in the tails. We set α to10% since we which to implement single-sided t-test with a 95% level of significance (i.e. 5% in one tail).
n is the number of scores (i.e. 12)
S_i is the sample standard deviation of the listener's 12 Diff Grade scores
$\overline{d_i}$ is the sample mean of the listener's 12 Diff Grade scores

If the statistic $T_i > 0$ for the listener *i*, then we conclude, with a 95% confidence, that the listener cannot reliably differentiate between the Hidden Reference and the System under Test, and the 12 listener responses are removed from consideration.

MUSHRA
Test 2, Test 3 and Test 4 use the MUSHRA test methodology. For each listener in each test, post-screening was based on listener scores for Hidden Reference and Low Pass filtered anchors. The procedure is as follows:

If, for any test item in a given test, either of the following criterion are not satisfied:

- The listener score for the hidden reference is greater than or equal to 90. That is HR >= 90.
- The listener scores the hidden reference, the 7.0 kHz lowpass anchor and the 3.5 kHz lowpass anchor are monotonically decreasing. That is, HR >= LP70 >= LP35.

Then all listener responses in that test are removed from consideration.

A.2 Statistical analysis

The statistical analysis of test scores follows standard statistical procedures. The calculation of the averages over the post-screened listener scores results in the Mean Subjective Score (MSS). The first analysis step of the results considers the calculation of the mean score $\bar{u}_{jk}$, for each of the presentations:

$$\bar{\mu}_{j,k} = \frac{1}{N}\sum_{i=1}^{N} \mu_{i,j,k}$$

where:

$\mu_{i,j,k}$ is the score of subject *i* for a given test condition *j* and test item *k*.
N is the number of subjects

Confidence intervals were derived from the standard deviation and the size of each sample. The 95% confidence interval for a given test condition *j* and test item *k* is given by:

$$[\bar{\mu}_{j,k} - \delta_{j,k}, \quad \bar{\mu}_{j,k} + \delta_{j,k}]$$

where

$$\delta_{j,k} = t_{\propto,n-1}\frac{S_{j,k}}{\sqrt{N}}$$

and the sample standard deviation $S_{j,k}$ is given by:

$$S_{j,k} = \sqrt{\sum_{i=1}^{N}\frac{(\bar{\mu}_{j,k} - \mu_{i,j,k})^2}{(N-1)}}$$

With a probability of 95%, the absolute value of the difference between the experimental or sample mean score and the "true" mean score (for a very high number of observers) is within the 95% confidence interval, on condition that the distribution of the individual scores are approximately Gaussian.

Similarly, a 95% confidence interval could be calculated for each test condition. In this case, sample means and sample standard deviations are calculated over all listeners and all test items.

Annex 4 Statistical analysis using ANOVA

Overview of ANOVA model

The objective of analysis of variance (ANOVA) is to assess whether a *treatment* applied to a set of samples has a significant effect, and to make that determination based on sound statistical principles [4], [5]. A treatment is, e.g., the processing of a signal by a coding system, but can also refer to other aspects of the experiment, so here we will to use the term *factor* instead of treatment.

The basic model of a score can be thought of as the sum of *effects*. A particular score may depend on which coding system was involved, which audio selection is being played, which laboratory is conducting the test, and which subject is listening. In other words, the score is the sum of a number of factor effects plus random error.

In terms of analyzing the data from the Verification Test, the following table lists the relevant factors in the experimental model. The test number (Test1, Test2, Test3, Test4) are not listed as factors since each test will be analyzed separately.

Factor	Description
Lab	Listening test site.
System	Coding system under test.
Signal	Test item.

The factors System and Signal form a fully-balanced and randomized factorial design, in that in every Test all Signals were processed by all Systems and were presented to the listeners for grading in random order. This balance has the advantage that the mean score for each system is an appropriate statistic for estimating the quality of that system.

The factors System and Signal are *fixed* in that they are specified in advance as opposed to being randomly drawn from some larger population.

Signal would be a *random* factor if the signals were actually selected at random from the population of all possible signals. Intuitively this is very appealing in that we might want to know how well the coding systems will perform for all possible speech and music items. However, we want the best coding system so the speech and music items were specifically selected because they are "difficult" items to code and so represent the "right tail" of a distribution of items rather than the entire population. Hence we have chosen to model Signal as a fixed factor.

The Labs, or test sites, was modeled as a random factor in that each Lab represents a specific test apparatus (i.e. listening room and audio presentation system) from a universe of possible test sites.

Since each Lab has a distinct set of listeners, the Listener factor is nested within the Labs factor. Listeners could be viewed as a random factor, in that it is intuitive and appealing to consider the listeners that participated in the test as representative of the larger population of all listeners. In this case the test outcome would represent the quality that would be perceived by the "typical" listener. However, the goal of the test was to have maximum discriminating capability so as to identify the best performing system. To this end, the subjects used were very experienced listeners that were "experts" at discerning the types of distortion typical of low-rate speech and audio coding. Regardless of these considerations, Listener was not used as a factor because of the very large number of levels (i.e. distinct listeners).

One aspect of the experimental design was not optimal, in that the Lab and Listener factors were not balanced. Participation as a test site and as a listener was voluntary, and a balanced design would have all sites and all listeners scoring all Tests, Systems and Signals, which was beyond the resources available within the MPEG Audio subgroup. However, the ANOVA calculations take the imbalance into account when computing the effects of each factor.

An important issue in using ANOVA is that it relies on several assumptions concerning the data set and the appropriateness of these assumptions should be checked as part of the data analysis. The most important assumptions are:

- The error has a Gaussian distribution.
- The variance of the error across factor levels is constant.

In addition, these assumptions must be valid to:

- Use parametric statistics for analysis of subjective data (which assumes that the error has a Gaussian distribution)
- Pool subjective data across test sites (which assumes that the variance of the error across test sites is constant)

Hence, aspects of ANOVA that validate these assumptions also validate the use of the statistical analysis used in the body of this report and described in Annex 3.

Finally, note that all ANOVA calculations, histogram and standard probability plots were performed using the R statistical package [6], [7].

Test 1

Test 1 uses the BS.1116 methodology, while Test 2, Test 3 and Test 4 use the MUSHRA test methodology. An ANOVA was done on the Diff Grades in Test 1, which made the data structure similar to that of Test 2. Hence, refer to explanations found in Section "Test 2," below, for an understanding of the meaning of the following tables and figures.

Model

Since there is only System under Test there is no factor "sys" in the ANOVA table.

```
             Df Sum Sq Mean Sq F value   Pr(>F)
lab           3   1.24  0.4133   2.726   0.0437 *
sig          11   9.81  0.8922   5.886 5.56e-09 ***
Residuals   465  70.48  0.1516
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

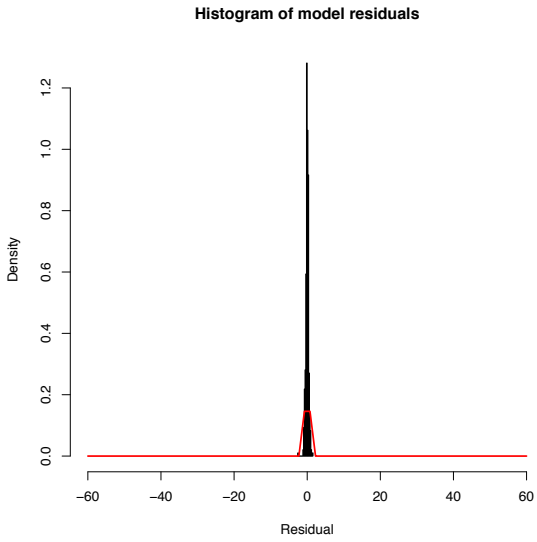
Performance

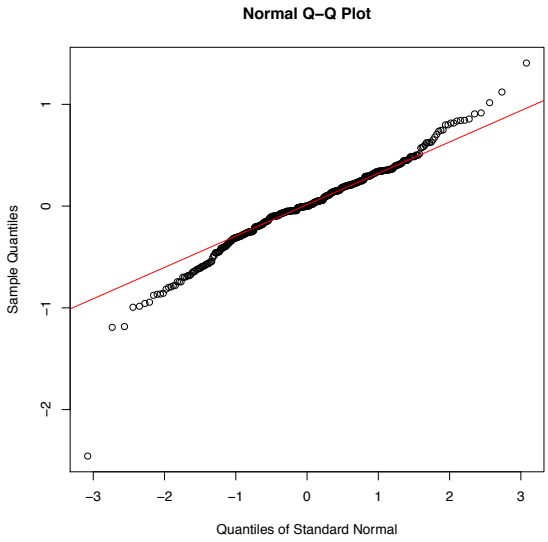
ANOVA CI is	0.035
Excel CI is	0.037

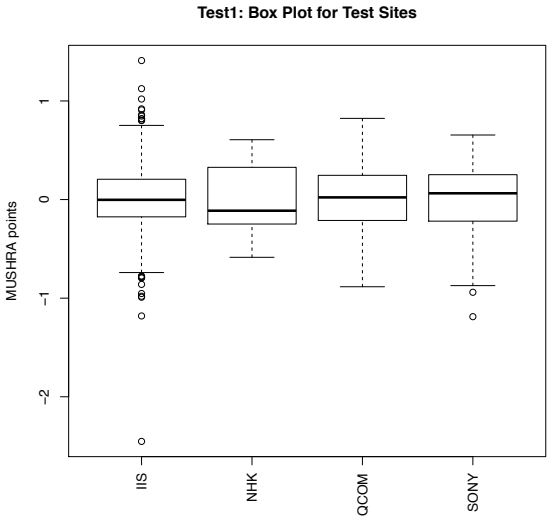
Verification of model assumptions

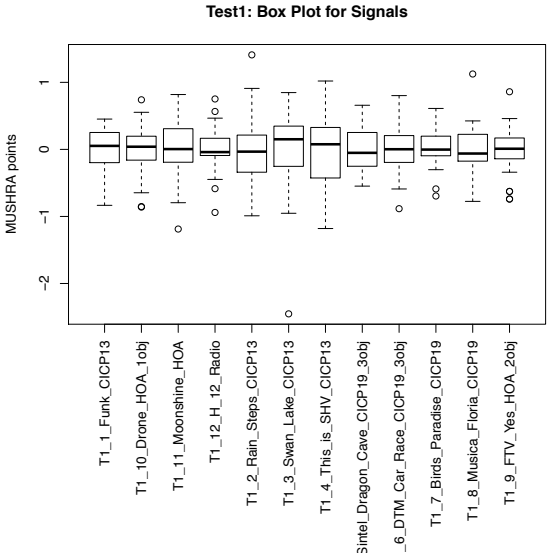
The histogram of the residual shows a very small range, but it is very close to having a Gaussian distribution, as shown in the Normal Q-Q plot. Hence used of parametric statistics is appropriate.

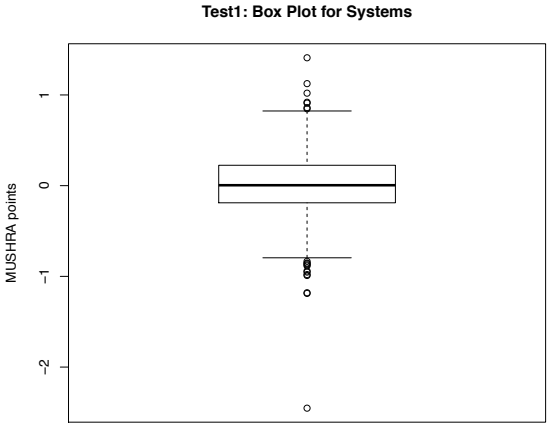
The box plot for Test Sites indicate that the residual variance is approximately the same for each value of the factor. Hence pooling of results from test labs is appropriate.


Histogram of model residuals
Density
Residual
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−60
−40
−20
0
20
40
60


Normal Q–Q Plot
Sample Quantiles
Quantiles of Standard Normal
−3
−2
−1
0
1
2
3


Test1: Box Plot for Test Sites
MUSHRA points
IIS
NHK
QCOM
SONY


Test1: Box Plot for Signals
MUSHRA points
T1_1_Funk_CICP13
T1_10_Drone_HOA_1obj
T1_11_Moonshine_HOA
T1_12_H_12_Radio
T1_2_Rain_Steps_CICP13
T1_3_Swan_Lake_CICP13
T1_4_This_is_SHV_CICP13
3intel_Dragon_Cave_CICP19_3obj
_6_DTM_Car_Race_CICP19_3obj
T1_7_Birds_Paradise_CICP19
T1_8_Musica_Floria_CICP19
T1_9_FTV_Yes_HOA_2obj


Test1: Box Plot for Systems
MUSHRA points

Test 2

Model

An aspect of ANOVA is to test the suitability of the model. A simple model incorporating all factors is expressed as:

```
Score = Lab + System + Signal + Error
```

The ANOVA report when using this model is:

```
              Df  Sum Sq Mean Sq  F value Pr(>F)
lab            2    1958     979   14.103  8e-07 ***
sig           11    1217     111    1.594 0.0936 .
sys            5 2837778  567556 8176.454 <2e-16 ***
Residuals   3077  213585      69
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The report indicates that model factors lab and sys are highly significant, while factor sig is not significant (at the 5% level of significance).

Performance

Using an ANOVA model does not change the mean score of the system under test. However, because it removes the factor mean effects from the error term, it reduces the error variance and hence the confidence interval on the mean scores. The CI Value (i.e. the ± value used to compute the 95% confidence interval) from ANOVA is

±0.720

In comparison, the average CI from grand mean analysis, as averaged over the systems under test, is

±0.746

Hence, we see that ANOVA gives slightly tighter confidence intervals.

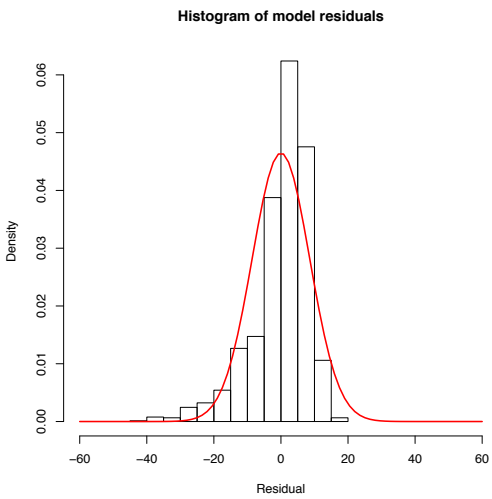
Verification of model assumptions

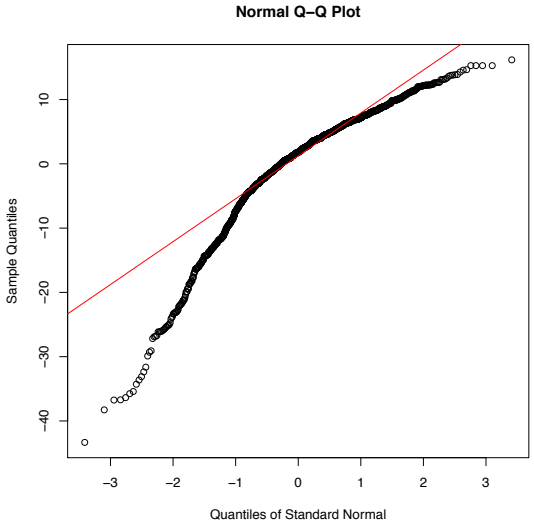
The following plots verify that the ANOVA residual has approximately a Gaussian distribution, as required for the validity of the ANOVA. Note that the systems Hidden Reference, 7.0 kHz low-pass original and 3.5 kHz low-pass original are removed prior to testing the ANOVA model assumptions since these systems do not get a truly random subjective assessment: subjects are instructed to score the Hidden Reference at 100 and generally tend to score the 7.0 kHz low-pass original and 3.5 kHz low-pass original as some nearly fixed score whose value is based on personal preference.

The left-hand plot below shows a histogram of the Test 2 residual with a best-fit Gaussian distribution (shown in red) superimposed on top. The right-hand plot shows a Normal Q-Q Plot of a Gaussian distribution (red line) and the Test1 residuals. The plot is

such that a true Gaussian distribution lies on a straight line. One can see that the Test1 residual deviates from the red line only at the ends, i.e. the tails of the distribution.

Both plots suggest that distribution of the Test 2 residuals are sufficiently close to a Gaussian distribution to apply parametric statistical analysis.

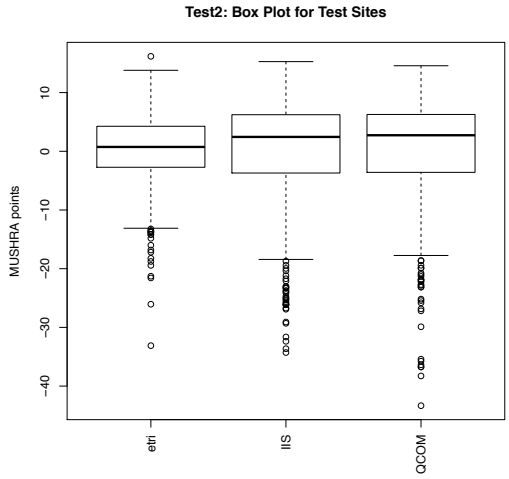


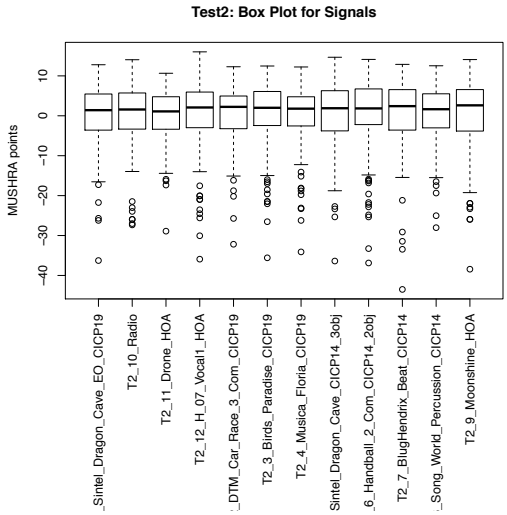


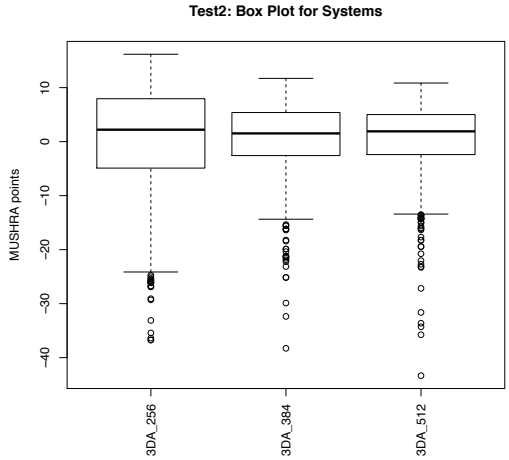
The following box plots show the scores associated with each level (or value) of the factors. For each of the factors Lab (Test Site), Test Item (Signals) and System under Test (System), the box plots indicate the distribution of score values after the factor effect is removed. In the box plots:

- The box indicates the range of the middle two quartiles of data (i.e. the box encompasses ±25% of the data, as measured from the mean).
- The "whiskers" indicate ±37.5% of the data, as measured from the mean
- The "circles" indicate data outliers that lie beyond of the ±37.5% region.

The plots indicate that the residuals have the approximately the same distribution for each value of the factor: Test Site and Signal spread is within a few tens of percent while System spread is within a factor of 2. Hence pooling of results from Labs is appropriate.







Test 3

The structure of Test 3 is similar to that of Test 2, so refer to explanations found in Section "Test 2," above, for an understanding of the meaning of the following tables and figures.

Model

```
                Df  Sum Sq Mean Sq  F value Pr(>F)
lab              2   26146   13073   92.530 <2e-16 ***
sig             11    3267     297    2.102 0.0173 *
sys              5 2800774  560155 3964.764 <2e-16 ***
Residuals     3149  444901     141
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

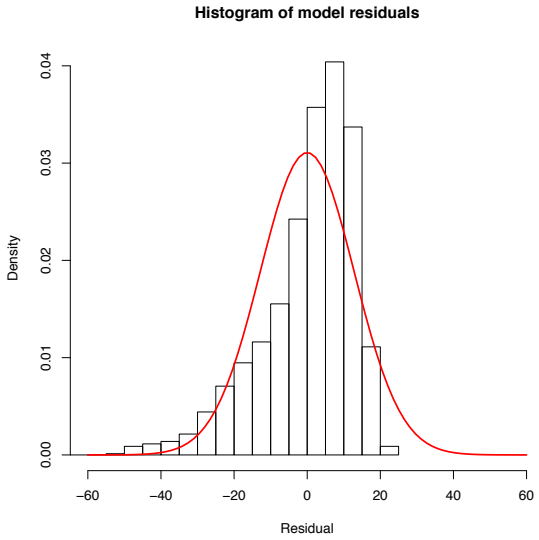
Performance

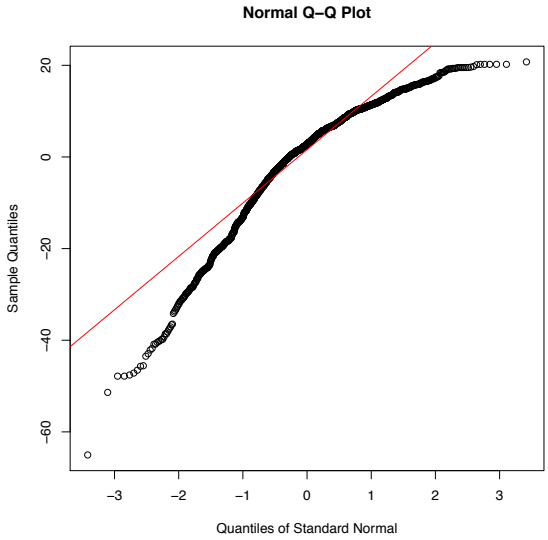
ANOVA CI is	±1.015
EXCEL CI is	±1.143

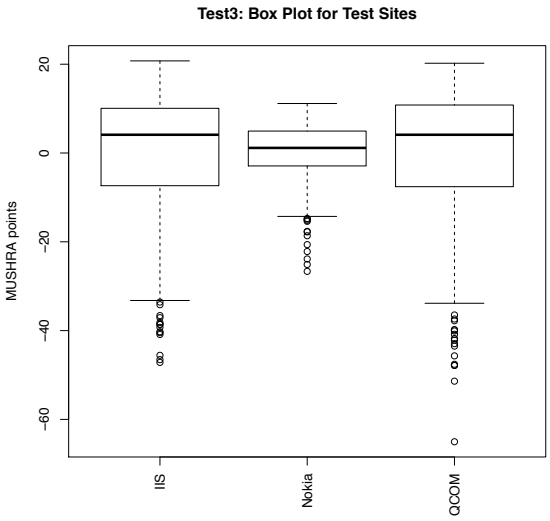
Verification of model assumptions

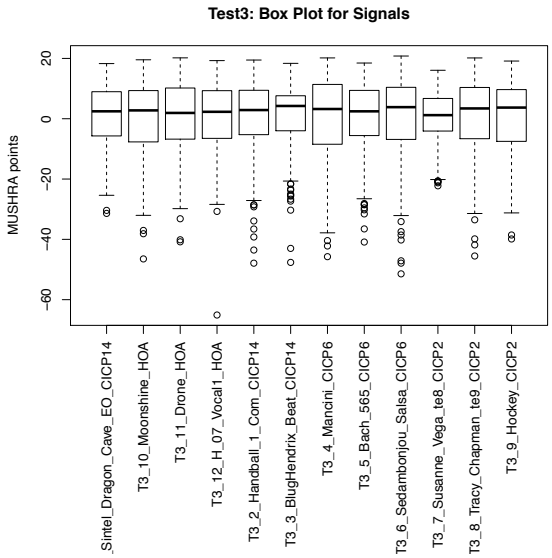
The histogram of the residual shows is close to having a Gaussian distribution, as shown in the Normal Q-Q plot. Hence it is appropriate to use parametric statistics.

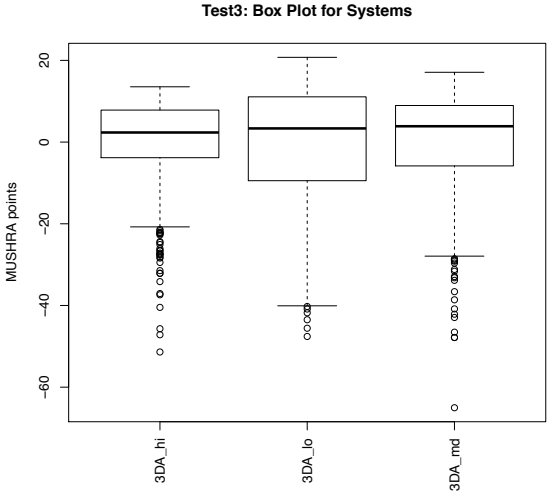
The box plot for Test Sites indicate that the residual variance is approximately the same (within a factor of 2 or 3) for each value of the factor. Hence pooling of results from Test Labs is appropriate.











Test 4

The structure of Test 4 is similar to that of Test 2, so refer to explanations found in Section "Test 2," above, for an understanding of the meaning of the following tables and figures.

Model

```
             Df  Sum Sq Mean Sq   F value    Pr(>F)
lab           4    5553    1388    15.490 1.48e-12 ***
sig          11    4997     454     5.068 6.68e-08 ***
sys           3 3610396 1203465 13427.473  < 2e-16 ***
Residuals  3245  290840      90
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

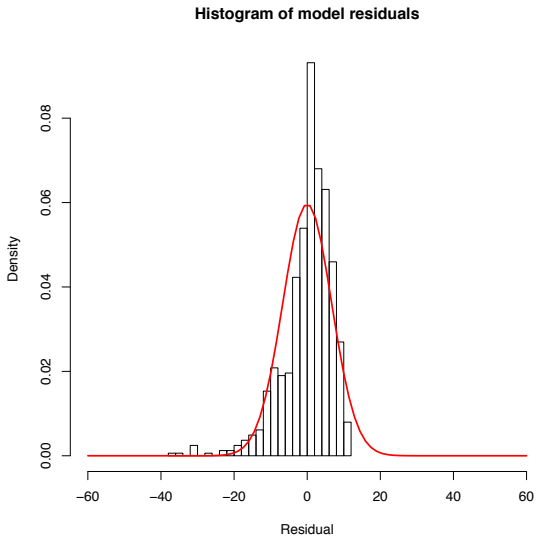
Performance

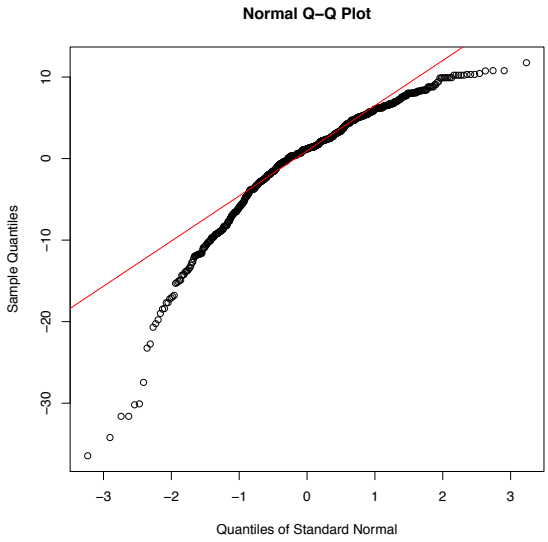
ANOVA CI is	0.650
EXCEL CI is	0.586

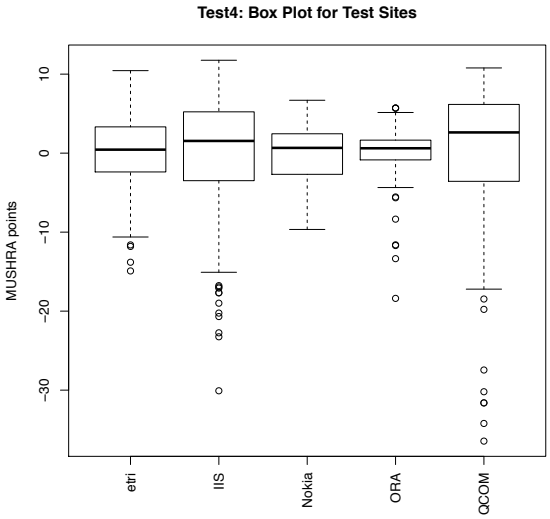
Verification of model assumptions

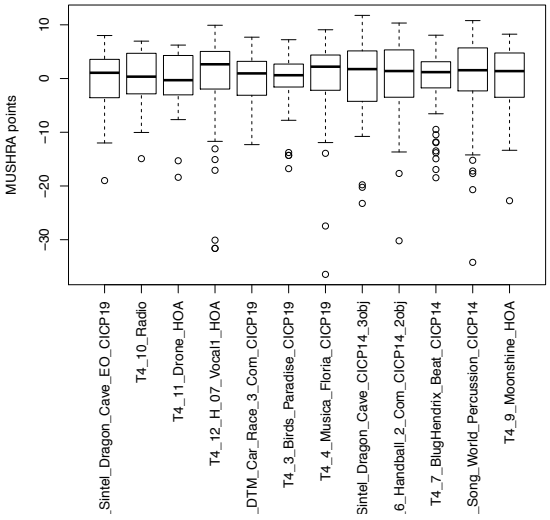
The histogram of the residual shows is close to having a Gaussian distribution, as shown in the Normal Q-Q plot. Hence it is appropriate to use parametric statistics.

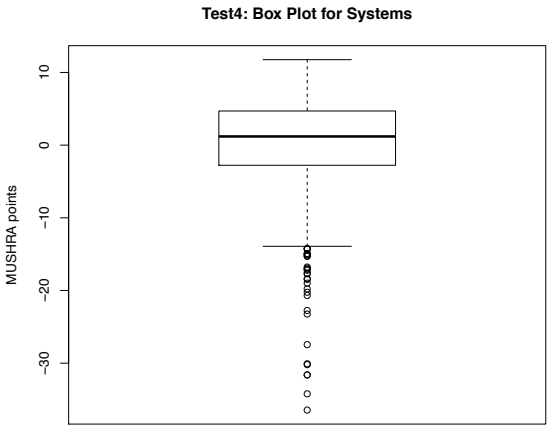
The box plot for Test Sites indicate that the residual variance is approximately the same (within a factor of 4) for each value of the factor. Hence pooling of results from Labs is appropriate.











References

[4] Montgomery, D.C. *Design and Analysis of Experiments*. John Wiley and Sons, New York, 1976.

[5] Bech, S. and Sacharov, N. *Perceptual Audio Evaluation, Theory, Method and Application*. John Wiley and Sons, Chinchester, West Sussex, England, 2002.

[6] Venables, W. N. and Ripley, D. B. *Modern Applied Statistics with S, Fourth Edition*. Springer, New York, 2002.

[7] The R Project for Statistical Computing, http://www.r-project.org/

Annex 5 Test item filenames

The tables below list the filename prefix for each test item in each of Test 1, Test 2 and Test 3. In each of these tests the test item is a set of mono signal files, one file per loudspeaker feed. If the prefix contains the string "CICPXX," then the "XX" indicates the loudspeaker presentation layout of the signal, where the "XX" is found in Table 1, below, under the heading "Channel Configuration."

The full filename for each loudspeaker signal associated with each item is constructed by appending the appropriate "A+XXX_E+YY" string found under the "Label" heading in Table 2, below, and finally adding the extension ".wav".

Test 1

Item	Filename
T1_1	T1_1_Funk_CICP13
T1_2	T1_2_Rain_Steps_CICP13
T1_3	T1_3_Swan_Lake_CICP13
T1_4	T1_4_This_is_SHV_CICP13
T1_5	T1_5_Sintel_Dragon_Cave_CICP19_3obj
T1_6	T1_6_DTM_Car_Race_CICP19_3obj
T1_7	T1_7_Birds_Paradise_CICP19
T1_8	T1_8_Musica_Floria_CICP19
T1_9	T1_9_FTV_Yes_HOA_2obj
T1_10	T1_10_Drone_HOA_1obj
T1_11	T1_11_Moonshine_HOA
T1_12	T1_12_H_12_Radio

Test 2

Item	Filename
T2_1	T2_1_Sintel_Dragon_Cave_EO_CICP19
T2_2	T2_2_DTM_Car_Race_3_Com_CICP19
T2_3	T2_3_Birds_Paradise_CICP19
T2_4	T2_4_Musica_Floria_CICP19
T2_5	T2_5_Sintel_Dragon_Cave_CICP14_3obj
T2_6	T2_6_Handball_2_Com_CICP14_2obj
T2_7	T2_7_BlugHendrix_Beat_CICP14
T2_8	T2_8_Song_World_Percussion_CICP14
T2_9	T2_9_Moonshine_HOA
T2_10	T2_10_Radio
T2_11	T2_11_Drone_HOA
T2_12	T2_12_H_07_Vocal1_HOA

Test 3

Item	Filename
T3_1	T3_1_Sintel_Dragon_Cave_EO_CICP14
T3_2	T3_2_Handball_1_Com_CICP14
T3_3	T3_3_BlugHendrix_Beat_CICP14
T3_4	T3_4_Mancini_CICP6
T3_5	T3_5_Bach_565_CICP6
T3_6	T3_6_Sedambonjou_Salsa_CICP6

T3_7	T3_7_Susanne_Vega_te8_CICP2
T3_8	T3_8_Tracy_Chapman_te9_CICP2
T3_9	T3_9_Hockey_CICP2
T3_10	T3_10_Moonshine_HOA
T3_11	T3_11_Drone_HOA
T3_12	T3_12_H_07_Vocal1_HOA

Test 4

For Test 4, the test item filename is constructed using the prefix in the Test 2 table and adding the suffix "binaural.wav." The Test 4 files are interleaved stereo WAV files, with interleave order L, R.

Table 1- Excerpt from ISO/IEC23008-3:2015, Table 95 ("Formats with corresponding number of channels and channel ordering")

Loudspeaker Layout Index or ChannelConfiguration as defined in ISO/IEC 23001-8	Number of Channels	Channels (with ordering)
2	2	CH_M_L030, CH_M_R030
6	6	CH_M_L030, CH_M_R030, CH_M_000, CH_LFE1, CH_M_L110, CH_M_R110
13	24	CH_M_L060, CH_M_R060, CH_M_000, CH_LFE2, CH_M_L135, CH_M_R135, CH_M_L030, CH_M_R030, CH_M_180, CH_LFE3, CH_M_L090, CH_M_R090, CH_U_L045, CH_U_R045, CH_U_000, CH_T_000, CH_U_L135, CH_U_R135, CH_U_L090, CH_U_R090, CH_U_180, CH_L_000, CH_L_L045, CH_L_R045
14	8	CH_M_L030, CH_M_R030, CH_M_000, CH_LFE1, CH_M_L110, CH_M_R110, CH_U_L030, CH_U_R030
19	12	CH_M_L030, CH_M_R030, CH_M_000, CH_LFE1, CH_M_L135, CH_M_R135, CH_M_L090, CH_M_R090, CH_U_L030, CH_U_R030, CH_U_L135, CH_U_R135

Table 2- Filename suffix for each presentation layout

No.	Label	Az °	El. °	2.0	5.1	5.1.2	7.1.4	22.2
1	A+000_E+00	0	0		X	X	X	X
2	A+030_E+00	30	0	X	X	X	X	X
3	A-030_E+00	-30	0	X	X	X	X	X
4	A+060_E+00	60	0					X
5	A-060_E+00	-60	0					X
6	A+090_E+00	90	0				X	X
7	A-090_E+00	-90	0				X	X
8	A+110_E+00	110	0		X	X		
9	A-110_E+00	-110	0		X	X		
10	A+135_E+00	135	0				X	X
11	A-135_E+00	-135	0				X	X
12	A+180_E+00	180	0					X
13	A+000_E+35	0	35					X
14	A+045_E+35	45	35					X
15	A-045_E+35	-45	35					X
16	A+030_E+35	30	35			X	X	
17	A-030_E+35	-30	35			X	X	
18	A+090_E+35	90	35					X
19	A-090_E+35	-90	35					X
20	A+110_E+35	110	35					
21	A-110_E+35	-110	35					
22	A+135_E+35	135	35				X	X
23	A-135_E+35	-135	35				X	X
24	A+180_E+35	180	35					X
25	A+000_E+90	0	90					X
26	A+000_E-15	0	-15					X
27	A+045_E-15	45	-15					X
28	A-045_E-15	-45	-15					X
29	LFE1_E-15	45	-15		X	X	X	X
30	LFE2_E-15	-45	-15					X

Annex 6 Listener Instructions

MPEG-H 3D Audio Verification Test
Test 1 – BS.1116 Methodology
Listener Instructions

Listeners must read these instructions and participate in the indicated training phase prior to their participation in the test phase.

Introduction

The MPEG Audio group has created a new standard for immersive audio coding, and this test will assess the audio quality that can be achieved by this technology under various operating conditions.

This listening test will use the so-called Double-Blind Triple Stimulus with Hidden Reference methodology.

Test procedure and User Interface

The figure below shows the graphical interface used for each trial to present one test item as processed by the systems under test. The buttons represent the reference (REF), which is always displayed at the bottom left, and all the systems to be graded, which are displayed as letter buttons "A" and "B". "REF" is always the reference (original) version of the audio item, against which both "A" and "B" are to be compared and graded. One of "A" or "B" is a processed version and the other is a hidden reference (identical to the reference). You are not told which of "A" and "B" is the processed version (hence the "blind" in the test name) and which is the hidden reference (hence the "hidden reference" in the test name). You will be able to switch freely among "REF", "A" and "B" at any time.

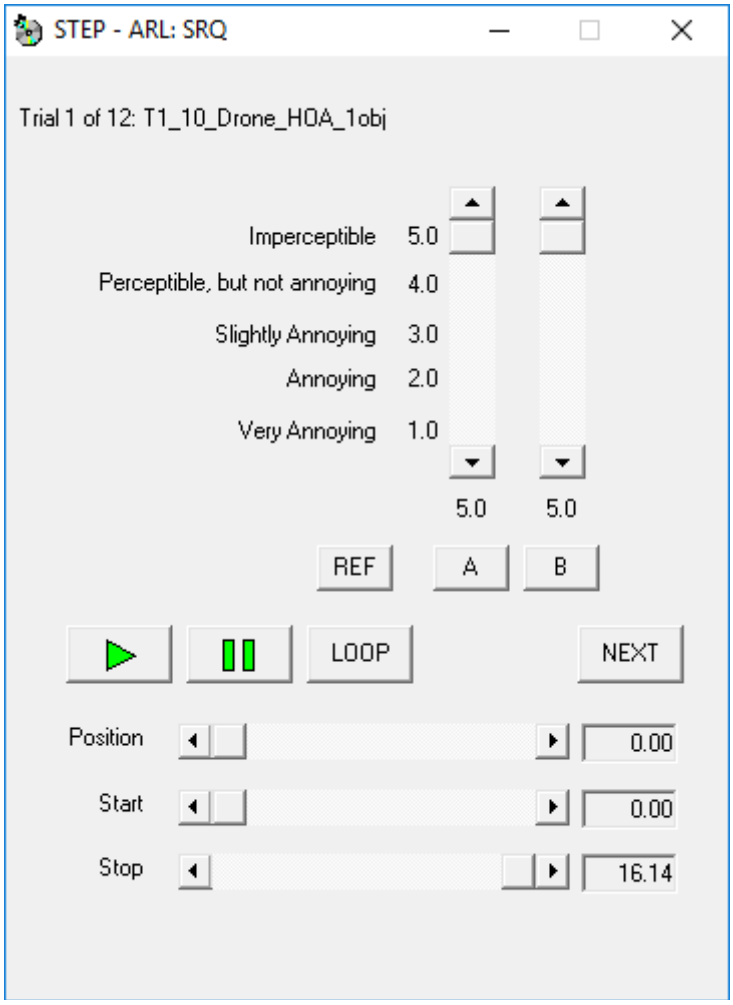
Above each button, with the exception of the button for the reference, a slider permits the listener to grade the quality of the systems under test on a continuous quality scale. The descriptors associated with the scale are

Imperceptible	(5.0)
Perceptible, but not annoying	(4.0)
Slightly annoying	(3.0)
Annoying	(2.0)
Very annoying	(1.0)

Note that any difference between the systems to be graded ("A" and "B") and the reference ("REF") shall be considered an impairment. Two grades must be given in each trial, one for "A" and one for "B". The grades serve two purposes:

- One grade must be 5.0, which is used to indicate which of "A" or "B" is the hidden reference.
- The other grade rates the difference between that item and the reference.

The trial number and the name of the test item are shown in the upper left of the graphical interface.



For each of the test items, the systems under test are randomly assigned to the letter buttons. In addition, the order of presenting the test items in the trails is randomized.

To begin the trial, the listener clicks on any button play audio. When another button is clicked, the audio presentation switches instantly and seamlessly from the one system to the other. Clicking on the "Loop" button plays the signal continuously. The horizontal Position slider indicates the instantaneous position in the signal waveform. Grabbing and moving the Start slider alters the start point for waveform looping, and similarly moving the Stop slider alters the end point, thus permitting a "loop and zoom" function that is particularly powerful for subjective evaluation. Rate the systems under test by grabbing and moving the vertical sliders above their corresponding letter buttons. When you are satisfied with the ratings, click on the "Next" button to go on to the next trial.

If the test is long and hence possibly fatiguing, you might want to interrupt the test and take a break after about 30 minutes. You can take a break after the completion of any trial. Please notify the test administrator if you choose to take a break.

When the last trial is scored, the Administrator window replaces the Trial window. Notify the test administrator that you have completed the listening session.

Training phase.

The purpose of the training phase is to allow listeners to identify and become familiar with potential distortions and artefacts produced by the test items. You will also become familiar with the test procedure and use of the test interface.

Please listen to the training signals to get a sense of how the processed signals sound relative to the reference signal. You should be considering during the training phase how you, as an individual, will interpret the audible impairments in terms of the grading scale, it is important that you should not discuss this personal interpretation with the other subjects at any time.

Test phase

The test phase will be carried out individually in test sessions each lasting about 30 to 60 minutes. In each trial, you will hear three versions, labelled "REF", "A" and "B" on the computer screen. "REF" is always the reference (original) signal against which both the "A" and "B" signals are to be compared and graded. One of "A" and "B" is a processed (coded/decoded) version and the other is a hidden reference (identical to the "REF" version).

You are asked to judge the "Overall Audio Quality" of the "A" and "B" versions in each trial. This attribute is related to any and all differences between the reference and the coded/decoded test item. Note that any difference between the reference and the coded/decoded item is to be considered as an impairment.

It is not possible to list all possible differences that may be created by the form of sound signal processing being evaluated in these tests. However what follows is a list of the main differences that may be expected.

It includes such things as harmonic distortions, added 'pops' or 'cracks', noise, temporal smearing, e.g. of sharp onsets, changes in loudness, changes in timbre, changes in spatial presentation, changes in background noise or reverberance. Anything else that the listener detects as a difference must be included in his/her overall rating.

In each trial, you are asked to rate the perceived difference (if any) between "REF" and "A" and the perceived difference between "REF" and "B" using the grading scale, which should be used as a continuous scale:

Imperceptible	(5.0)
Perceptible, but not annoying	(4.0)
Slightly annoying	(3.0)
Annoying	(2.0)
Very annoying	(1.0)

Note that any difference between the systems under test ("A", "B", etc.) and the reference ("REF") shall be considered an impairment. Two grades must be given in each trial, one for "A" and one for "B". The grades serve two purposes:

- One grade must be 5.0, which is used to indicate which of "A" or "B" is the hidden reference.
- The other grade rates the difference between that item and the reference.

MPEG-H 3D Audio Verification Test
Test 2, 3, 4 – MUSHRA Methodology
Listener Instructions

Listeners must read these instructions and participate in the indicated training phase prior to their participation in the test phase.

Introduction

The MPEG Audio group has created a new standard for immersive audio coding systems, and this test will assess the audio quality that can be achieved by this technology under various operating conditions.

This listening test will use the MUSHRA test methodology, which has the advantage of displaying all stimuli (both coding systems and anchor systems) for a given test item. Hence you are able to directly compare the stimuli in the course of giving a grade to each.

Test Procedure and User Interface

The figure below shows the graphical interface used for each trial to present one test item as processed by all systems under test. The buttons represent the reference (REF), which is always displayed at the bottom left, and all the systems to be graded, including the codecs under test, reference codecs, hidden reference and anchor signals (band-limited processed references), which are displayed as letter buttons. “REF” is always the reference (original) version of the audio item, against which the letter systems (“A”, “B”, etc.) are to be compared and graded.

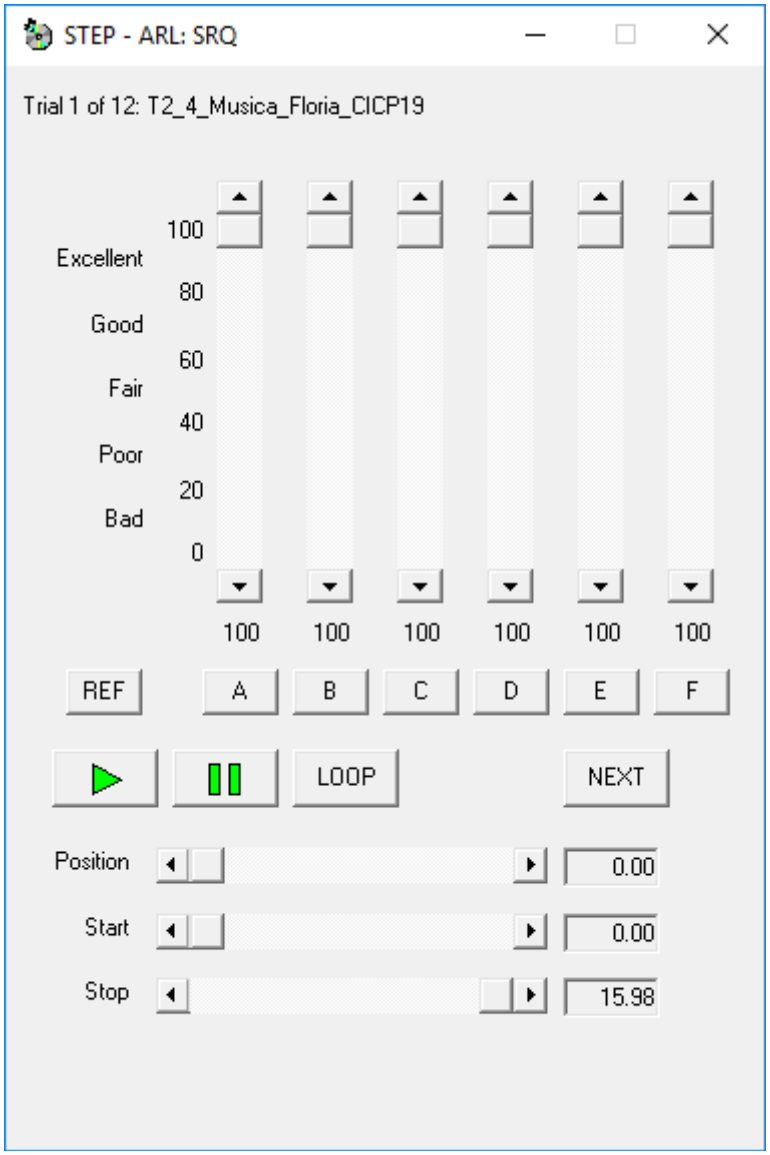
Above each button, with the exception of the button for the reference, a slider permits the listener to grade the quality of the systems under test on a continuous quality scale. The descriptors associated with the scale are

- Excellent (80-100).
- Good (60-80)
- fair (40-60)
- poor (20-40)
- bad (0-20)

Note that any difference between the systems under test (“A”, “B”, etc.) and the reference (“REF”) shall be considered an impairment. When assigning grades in each trial:

- One grade must be 100, which is used to indicate the hidden reference.
- The other grades rate the difference between that item and the reference.

The trial number and the name of the test item are shown in the upper left of the graphical interface.



For each of the test items, the systems under test are randomly assigned to the letter buttons. In addition, the order of presenting the test items in the trails is randomized.

To begin the trial, the listener clicks on any button play audio. When another button is clicked, the audio presentation switches instantly and seamlessly from the one system to the other. Clicking on the "Loop" button plays the signal continuously. The horizontal Position slider indicates the instantaneous position in the signal waveform. Grabbing and moving the Start slider alters the start point for waveform looping, and similarly moving the Stop slider alters the end point, thus permitting a "loop and zoom" function that is particularly powerful for subjective evaluation. Rate the systems under test by grabbing and moving the vertical sliders above their corresponding letter buttons. When you are satisfied with the ratings, click on the "Next" button to go on to the next trial.

If the test is long and hence possibly fatiguing, you might want to interrupt the test and take a break after about 30 minutes. You can take a break after the completion of any trial. Please notify the test administrator if you choose to take a break.

When the last trial is scored, the Administrator window replaces the Trial window. Notify the test administrator that you have completed the listening session.

Training phase
The purpose of the training phase is to allow listeners to identify and become familiar with potential distortions and artefacts produced by the test items. You will also become familiar with the test procedure and use of the test interface.

Please listen to the training signals to get a sense of how the processed signals sound relative to the reference signal. You should be considering during the training phase how you, as an individual, will interpret the audible impairments in terms of the grading scale, it is important that you should not discuss this personal interpretation with the other subjects at any time.

Test phase
The test phase will be carried out individually in test sessions each lasting about 30 to 60 minutes. In each trial, you will hear several versions of the test items. The "REF", buttons is the reference (original) signal, and the letters "A", "B", etc. are associated with a different version of the signal, i.e. the original processed by one of the systems under test.

You are asked to judge the "Overall Sound Quality" of the versions of the test item in each trial. This attribute is related to any and all differences between the reference and the coded/decoded test item. Note that any difference between the reference and the coded/decoded item is to be considered as an impairment.

It is not possible to list all possible differences that may be created by the form of sound signal processing being evaluated in these tests. However what follows is a list of the main differences that may be expected.

It includes such things as harmonic distortions, added 'pops' or 'cracks', noise, temporal smearing, e.g. of sharp onsets, changes in loudness, changes in timbre, changes in spatial presentation, changes in background noise or reverberance. Anything else that the listener detects as a difference must be included in his/her overall rating.

In each trial, you are asked to rate the perceived difference (if any) between "REF" and of the systems under test ("A", "B", etc.) using the following grading scale, which should be used as a continuous scale:

Excellent	(80-100)
Good	(60-80)
Fair	(40-60)
Poor	(20-40)
Pad	(0-20)

Note that any difference between the systems under test ("A", "B", etc.) and the reference ("REF") shall be considered an impairment. When assigning grades in each trial:

- One grade must be 100, which is used to indicate the hidden reference.
- The other grades rate the difference between that item and the reference.

Test 4 – Headphone Listening
The stimuli in Test 4 are presented via headphones, but are intended to have the same spatial resolution as tests presented via loudspeakers.