TR 043

‘EAR’ LISTENING TEST REPORT USING THE MULTIPLE STIMULUS IDEAL PROFILE METHOD

TESTING THE EBU ADM RENDERER

TECHNICAL REPORT

Geneva
October 2018
Contents

1. INTRODUCTION .............................................................................................................. 5
2. AIMS OF THE EBU-BTF STUDY .................................................................................. 5
3. SUMMARY DESCRIPTION OF THE MS-IPM METHOD ............................................... 6
4. DESCRIPTION OF THE TEST ......................................................................................... 6
   4.1 EXPERIMENTAL PARAMETERS .............................................................................. 6
      4.1.1 Laboratories ...................................................................................................... 6
      4.1.2 Assessor pre-selection ....................................................................................... 7
      4.1.3 Speaker layouts ............................................................................................... 7
      4.1.4 Renderers ......................................................................................................... 7
      4.1.5 Samples selection ............................................................................................ 8
      4.1.6 Test conditions ................................................................................................ 8
      4.1.7 Stimulus preparation ...................................................................................... 9
    4.2 DATA COLLECTION .................................................................................................. 9
      4.2.1 Quality scaling ................................................................................................. 10
      4.2.2 Attribute selection ......................................................................................... 10
      4.2.3 Ideal profile rating .......................................................................................... 10
   4.3 EXPERIMENTAL DESIGN ......................................................................................... 10
      4.3.1 Blocking and stimulus presentation ................................................................. 10
      4.3.2 User interface ................................................................................................. 12
      4.3.3 Summary of test administration details .......................................................... 12
5. DATA ANALYSIS ......................................................................................................... 14
   5.1 HIGH LEVEL OVERVIEW ..................................................................................... 14
      5.1.1 Overall Means and CI (BAQ only) ................................................................. 14
      5.1.2 Raw attribute data (spider plots) ..................................................................... 14
      5.1.3 Multivariate analysis ...................................................................................... 15
   5.2 DETAILED ANALYSIS ........................................................................................... 17
      5.2.1 Average test condition performance per attribute ............................................ 17
      5.2.2 Impact of samples on system performance ...................................................... 18
   5.3 DATA QUALITY ...................................................................................................... 19
      5.3.1 Assessor screening .......................................................................................... 19
      5.3.2 Repeatability within lab .................................................................................. 20
      5.3.3 Repeatability per test condition .................................................................... 21
6. SUMMARY OF RESULTS ........................................................................................... 23
7 CONCLUSIONS ............................................................................................................. 24
8 ACKNOWLEDGEMENTS ............................................................................................... 24
9 REFERENCES ................................................................................................................ 24

ANNEX 1: OBSERVATIONS FOR IMPROVEMENT OF THE MS-IPM ...................................... 27
EBU ADM Renderer listening test report
Using the Multiple Stimulus Ideal Profiling Method (MS-IPM)

<table>
<thead>
<tr>
<th>EBU Committee</th>
<th>First Issued</th>
<th>Revised</th>
<th>Re-issued</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>2018</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Keywords: EAR, EBU ADR Renderer, Multiple Stimulus Ideal Profiling Method MS-IPM, Ideal Profiling, audio subjective assessment, BTF.

1. Introduction
The EBU has published EBU Tech 3388 [1], which is a specification of a native ADM [2] audio renderer intended for use by broadcasters in the production and monitoring of next generation audio (NGA) workflows. The EAR is one of the systems assessed in this test.

The purpose of the listening test was to evaluate the performance of audio renderers in a range of broadcast and standard listening rooms, with different loudspeaker channel configurations. An Multiple Stimulus - Ideal Profile Method (MS-IPM) [3] experiment was designed to compare 7 test conditions with 6 programme items. The test conditions comprised a combination of channel layouts (0+2+0, 0+5+0, 4+7+0, 9+10+3), in accordance to ITU-R BS.2051-1 [4], two different renderers and one down-mix. The identical double-blind test design was performed in 5 different laboratories, comprising of either ITU-R BS.1116-3 [5] compliant listening rooms or broadcast listening labs.

2. Aims of the EBU-BTF study
The aim of this study was primarily to evaluate and compare the performance of different renderers with broadcast content, with different speaker configurations, in comparable listening conditions. The previous studies driven by EBU members regarding spatialized sound had raised the necessity to take several perceptive dimensions into account to better understand how a listener rates the audio quality. This is why this study digs deeper in the perceptual attribute characteristics beyond the basic audio quality. Also due to the nature of rendered audio, a predefined reference is not evident, such the applications of test methodologies such as ITU-R BS.1116 Error! Reference source not found. or ITU-R BS.1534 [6] are not possible. The MS-IPM method was chosen as a non-reference method, furthermore allowing for the comparative analysis based on overall quality and also attribute characteristics. Typical broadcast productions involving spatial sound production and representing various genres have been selected by EBU members. 5 calibrated listening rooms of various sizes and broadcast quality equipment were employed.

The evaluation sought to study
- The overall performance of renderers with different loudspeaker configurations
- The impact of sound samples on renderer performance
- The overall quality of renderers
- An in-depth analysis of renderer performance using sound quality attributes
The evaluation did not aim to study
- Individual assessor performance, beyond post screening
- Individual laboratory results

3. **Summary description of the MS-IPM method**

The MS-IPM method was developed as an extension of the IPM method employed in the development of consumer products [14] [15] and was originally employed for the evaluation of hearing aids [12]. This successful application in audio illustrated the power of the method to evaluate the performance of technology from multiple perspectives, including the overall quality, but also how the technology is perceived in a more detailed manner, using attribute-based characterisation.

The MS-IPM uses the multiple stimulus presentation approach employed in Recommendation ITU-R BS.1534 [6] as a basis for comparison of the test conditions under test. The assessor is asked to provide ratings based on:
- Overall subjective quality
- Attribute rating (predefined sets of selected attributes)
- Ideal profile ratings (per attribute)

The overall subjective quality ratings are performed using the Continuous Quality Scale as defined in ITU-R BS.1534 [6]. The next stage is attribute rating, as found in ITU-T P.835 [20] and ITU-T P.806 [21]. Pertinent sound quality attributes, ideally from existing and validated lexicons (e.g. as found in [8] [9] [10] & [11]), were pre-selected by a panel of experts based on the characteristics rendered sound samples. Using these attributes assessors provide their ratings on 100-point line scales with end-point labels and run-offs. As part of the attribute rating stage, the assessor is asked to provide a rating for an envisaged ideal system on each attribute scale - called the ideal rating. In this way, an estimate of the ideal system profile can be gained from assessors, to complement the overall subjective quality and attribute ratings of the test conditions within the experiment.

4. **Description of the test**

During the test planning, it was desired that the test be performed in several different laboratories. Assessors only performed the test in their own laboratory. Furthermore, to accommodate this aspect, the attribute and instructions were translated into the local languages of each laboratory, the nature of which is described in Table 1. In all other regards an identical test was setup for each laboratory, such that the data could be pooled for analysis.

The test comprised of 8 test conditions combining renderers and speaker layouts.

4.1 **Experimental parameters**

4.1.1 **Laboratories**

Table 1 provides a summary of the 5 listening room employed for this test.
Table 1: Summary of the listening room characteristics for each laboratory.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Lab 1</th>
<th>Lab 2</th>
<th>Lab 3</th>
<th>Lab 4</th>
<th>Lab 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0+2+0 ITU-R BS.2051 compliant</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>0+5+0 ITU-R BS.2051 compliant</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>4+7+0 ITU-R BS.2051 compliant</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>9+10+3 ITU-R BS.2051 compliant</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Speaker type</td>
<td>Nubert nu8Box 101 (driven by Samson Servo 120a amplifiers)</td>
<td>Genelec 8030B</td>
<td>Passive speakers based on Km205f from DieProduction driven by RAM audio amplifiers</td>
<td>Musikelectronic Geithain RL906</td>
<td>Genelec 8320</td>
</tr>
<tr>
<td>ITU-R BS.1116-3 compliant</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Dimensions (W x L x H), m</td>
<td>6.78 x 5.27 x 3.21</td>
<td>5.05 x 5.63 x 2.88</td>
<td>4.87 x 5.95 x 2.5</td>
<td>6.00 x 7.96 x 4.50</td>
<td>4.64 x 7.84 x 2.62</td>
</tr>
</tbody>
</table>

### 4.1.2 Assessor pre-selection

In each laboratory, assessors were recruited from technical staff, where at all possible with experience in performing listening tests. Further in-depth screening was not performed. A total of 58 assessors performed the experiment.

### 4.1.3 Speaker layouts

To evaluate the performance of the different rendering technology in different configurations, four loudspeaker layouts were selected, all based upon ITU-R BS.2051-1 [4], including:

- 0+2+0
- 0+5+0
- 4+7+0
- 9+10+3

All laboratories were setup with these configurations within the tolerance limits defined in ITU-R BS.2051-1 [4]. For listening rooms complying with ITU-R BS.1116-1 [5], the speakers where equalised and calibrated to meet the operational room response characteristics.

The speaker setup was level calibrated in an identical manner in each laboratory using a common calibration sample with the following characteristics: 24 channel, 48 kHz, 24-bit pink noise file @ -18 dBFS. The calibration levels at the listening position in all listening rooms was set to 72 ± 0.25 dBA and ensured a comfortable listening level with the test samples.

### 4.1.4 Renderers

The test was performed using several systems comprising 2 renderers (Sys1 - Sys2) and a downmix (Sys3), defined as follows:

- **Sys1** represented the EBU ADM Renderer, specified in EBU Tech 3388 [1];
- **Sys2** was a commercially available renderer which was integrated in a Digital Audio Workstation, capable of processing ADM input files;
- **Sys3** consisted of a downmix of the 9+10+3 rendering of Sys2 to a 4+7+0 loudspeaker layout. The downmix was performed using the matrix defined in § 4.1.7.
4.1.5 Samples selection

Three samples were taken from the S3A\(^1\) object-based audio drama dataset [16]. These are spatial audio drama scenes, available in the ADM BW64 format. All scene components are defined as audio objects with 3D position meta-data, many of which are time varying, and sources are placed at directions all around the listener including at elevated positions. They were mixed at BBC R&D on a 32-loudspeaker system with two subwoofers, using a VBAP-based 3D-panner. They were designed as material suitable for demonstration and testing of object-based spatial audio systems.

The Protest clip featured a dense crowd scene, set outdoors, with many voices spread around the listener at a range of elevations. There was distinct foreground dialogue, as well as background crowd voices. A helicopter flies through the scene from rear-left to front-right and over the listener’s head. It also features a 16-channel 3D reverb bus. The clip was 27s long.

The Forest clip featured a forest ambience, comprised of a 3D 16-channel field recording with widely spaced cardioid microphones in two height layers, as well as a number of sound effects. A stereo location recording of children running and laughing is automated to pan from rear-right to rear-left and then around to front-left. There is also music in the clip, with instruments spread around the scene. The clip was 23s long.

The Family clip featured a domestic interior scene with three family members talking around the listener, from front-left to the right. There is some movement, with accompanying footsteps. There is also another character directly overhead, with the impression of muffled shouting and movement of heavy objects coming through the ceiling. The scene features room ambience as well as exterior road noise and a distant washing machine sound. It also uses a 16-channel 3D reverb. The clip was 26s long.

For further details on the creation of these items see [17].

Four samples were excerpted from different object-based audio productions of France Télévisions and Radio France.

The Sequences clip featured a live musical session where a DJ plays several acoustic and electronic sounds simultaneously which are spatially mixed using an object-based audio workflow. This excerpt comes from one of the thirty 45 minutes sessions produced by Radio France, Milgram and France Télévisions.

The 4everShort clip was excerpted from a short movie - 3 minutes in total - produced by France TV innovations & developments to study a complete Ultra High Definition workflow including object-based audio. The sound track has been mixed in 3D by using a combination of Ambisonics ambience, MS and lavalier microphones with several 2D reverberations.

The Monteverdi clip was excerpted from a 3D recording of a classical music piece produced by Wahoo Production, Chateau de Versailles Spectacles and France Télévisions. The audio track of this production has been mixed in 7+9+0.

The Roland Garros clip was excerpted from the men finale of the 2017 edition which was recorded in multitrack (around 30 microphones signals) and live mixed in Dolby Atmos. A specific remix of the chosen excerpt was produced in 3D thanks to an object-based audio workflow.

4.1.6 Test conditions

The test conditions comprised of a combination of the renderers (Sys1...Sys3) and speaker layouts (0+2+0, 0+5+0, 4+7+0, 9+10+3). A total a test conditions were available for evaluation, which are summarised as follows:

---

\(^1\) S3A: Future Spatial Audio for an Immersive Listener Experience at Home [http://www.s3a-spatialaudio.org/](http://www.s3a-spatialaudio.org/)
4.1.7 Stimulus preparation

Each item was made in or imported into the production tools for Sys2, which were used to render the unprocessed stimuli for the conditions Sys2-0-2-0, Sys2-0-5-0 and Sys2-4-7-0.

An ADM-BWF file for each item was exported from the production tools for Sys2. These were rendered using the EAR to produce the unprocessed stimuli for conditions Sys1-0-2-0, Sys1-0-5-0, Sys1-4-7-0 and Sys1-9-10-3.

The following downmix matrix was applied to the unprocessed stimuli for condition Sys1-9-10-3 to produce the unprocessed stimuli for condition Sys3-4-7-0:

- \( M+000 = B+000 + M+000 \)
- \( M+030 = B+045 + M+030 + 0.707107 \times M+060 \)
- \( M+090 = 0.707107 \times M+060 + M+090 \)
- \( M+135 = M+135 + 0.707107 \times M+180 \)
- \( M-030 = B-045 + M-030 + 0.707107 \times M-060 \)
- \( M-090 = 0.707107 \times M-060 + M-090 \)
- \( M-135 = 0.707107 \times M+180 + M-135 \)
- \( U+045 = 0.5 \times T+000 + 0.707107 \times U+000 + U+045 + 0.707107 \times U+090 \)
- \( U+135 = 0.5 \times T+000 + 0.707107 \times U+090 + U+135 + 0.707107 \times U+180 \)
- \( U-045 = 0.5 \times T+000 + 0.707107 \times U+000 + U-045 + 0.707107 \times U-090 \)
- \( U-135 = 0.5 \times T+000 + 0.707107 \times U+180 + 0.707107 \times U-090 + U-135 \)

The unprocessed renderings for each item were then time-aligned to maximise the correlation between the conditions, trimmed to remove silence at the start and end, and level-normalised to -23 LKFS according to EBU R128 [19] with no per-channel weighting.

These aligned and normalized stimuli were then loudness matched by ear [insert details] to produce the final stimuli used in the test.

4.2 Data collection

The test was built and administered using SenseLabOnline.com, allowing for identical tests to be performed in all laboratories in the local language. For each test, assessors were first instructed verbally and in writing. They then performed a short familiarisation and training with the user interface, enabling them to listen to all the audio samples and become familiar with the attributes. More extensive training of assessors was not possible within the scope of this test. Once the familiarisation was completed, the assessors progressed onto the test, which was performed over multiple sessions over multiple days, with an average total duration of ~4 hours. Assessors where prompted by the software to take breaks approximately every 20 minutes.
4.2.1 Quality scaling
Overall quality was evaluated using an overall quality scale as found in ITU-R BS.1534, comprising of a 100-point scale as illustrated in Figure 1.

4.2.2 Attribute selection
Attributes where adopted from two existing lexicons as provided in ITU-R Report BS.2399 [8] and Spatial Audio Quality Inventory (SAQI) [9]. Experts from each laboratory listened to the test samples during a test pilot phase and selected the most pertinent attributes, from which the most commonly occurring 5 attributes were chosen for inclusion in the test.

The 5 attributes were then translated from English into 4 other languages, as presented in Table 2.

4.2.3 Ideal profile rating
The ideal rating and associated ideal profile are new concepts in our field and thus worth a few words. The ideal rating is provided by assessor for each attribute, using the same scale that was used to rate each of the test conditions. The difference with the ideal rating is that no test condition is present. The assessor thus needs to consider what is the desired level of this attribute for a hypothetical ideal system they can envisage based on their experience and the test conditions they have heard during the trial.

The ideal profile then is calculated at the analysis phase and comprises of the average assessor scores for each test condition for each attribute. This profile can be presented averages over all test samples or individually for each test sample. The ideal profile is an indication of the assessor’s expectations, irrespective of technological limitations and can be used as a frame of references to compliment the overall quality scores in the evaluation of the performance of test conditions.

4.3 Experimental design

4.3.1 Blocking and stimulus presentation
An experimental design was developed to answer the basic research questions. The aim was to ensure that an identical experiment could be performed efficiently in all 5 laboratories and be completed in approximately 4 hours - estimated to be about twice as long as a ITU-R BS.1116-3 [5] listening test.

Due to the fact that the renderers and downmix were not available for all speaker layouts, the 8 test conditions were developed as a combination of the test systems and samples, as already described in § 4.1.6.

A full factorial design was created that could be administered in each laboratory, whereby all combination of test conditions was processed for all samples. As assessors would only perform the test at their own laboratory, the assessors could be considered as nested within laboratory.

For each sample, assessors were first asked to evaluate the overall subjective quality for each test condition. Thereafter the 5-attributes were to be rated for each test condition. In all cases, test conditions were randomly presented for each screen/trial. For each attribute the assessors were also asked to provide their ideal rating. The order of attributes was randomised for each assessor.

Additionally, two of the samples (Forest and Monteverdi) were replicated in order to study assessor performance and replication of the experiments within each lab and overall. These replications were handled as additional test samples and randomly presented to the assessors during the test sequence.
<table>
<thead>
<tr>
<th>Attribute name</th>
<th>Definition</th>
<th>Attribute name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Scene depth</td>
<td>The radial extent of scene, or ensemble from the listener (in any direction away from the listener).</td>
<td>4. Color tone</td>
<td>Temporal impression which is determined by the ratio of high to low frequency components.</td>
</tr>
<tr>
<td>2. Profondeur de la scène</td>
<td>Scale: Shallow-Deep</td>
<td>5. Clarté</td>
<td>Sensation of a ton drug/paste or in graves or en graves by example, or occurrence sensation of a ton or metallic</td>
</tr>
<tr>
<td>8. Japanese</td>
<td>Sensation of a ton drug/paste or in graves or en graves by example, or occurrence sensation of a ton or metallic</td>
<td>9. Japanesse</td>
<td>Klangfarbe farbe</td>
</tr>
<tr>
<td>9. Japanese</td>
<td>Sensation of a ton drug/paste or in graves or en graves by example, or occurrence sensation of a ton or metallic</td>
<td>10. Japanesse</td>
<td>Klangfarbe farbe</td>
</tr>
<tr>
<td>11. Danish</td>
<td>Sensation of a ton drug/paste or in graves or en graves by example, or occurrence sensation of a ton or metallic</td>
<td>12. Danises</td>
<td>Klangfarbe farbe</td>
</tr>
<tr>
<td>13. German</td>
<td>Sensation of a ton drug/paste or in graves or en graves by example, or occurrence sensation of a ton or metallic</td>
<td>14. Sachsen</td>
<td>Klangfarbe farbe</td>
</tr>
<tr>
<td>15. English</td>
<td>Sensation of a ton drug/paste or in graves or en graves by example, or occurrence sensation of a ton or metallic</td>
<td>16. England</td>
<td>Klangfarbe farbe</td>
</tr>
<tr>
<td>17. French</td>
<td>Sensation of a ton drug/paste or in graves or en graves by example, or occurrence sensation of a ton or metallic</td>
<td>18. France</td>
<td>Klangfarbe farbe</td>
</tr>
</tbody>
</table>
4.3.2 User interface

**Figure 1:** Graphical user interface of the overall subjective quality rating phase.

**Figure 2:** Graphical user interface of the attribute rating phase, illustrated with the attribute “clarity”. The ideal rating for this attribute is provide using the right hand scale, labelled with an “***”.

4.3.3 Summary of test administration details

The test was performed in an identical manner in each of the laboratories, using the same instructions, translated into the local language. The test was performed by one assessor at a time. Each assessor was initially provided with written and verbal instruction on the test protocol and the attribute definitions. They were then allowed to perform a familiarisation test, whereby all samples were presented using a multiple stimulus presentation for one of the attributes. Assessors were encouraged to spend as much time as possible to become familiar with the samples and the attributes. Upon completion of the familiarisation phase, the assessors were encouraged to ask any question about the test, user interface, attribute definitions, etc.
Following a short break, assessors moved onto the main test by re-reading the attribute definitions and the following test instructions:

<table>
<thead>
<tr>
<th>Once you have completed the familiarisation, please take a break. Have a read of the attribute definitions again and clarify anything that is unclear with your test administrator.</th>
</tr>
</thead>
</table>

When you are ready you can start the **Main test** fr in SenseLabOnline. You will have 48 screens/trails to complete in total in about 3-4 hrs (including time for familiarization). Please take breaks when you need. We encourage you to do this over more than one day. You can close the test at any time and when you login again, the test will resume where you last left off.

During the test, you will be asked to evaluate the **Basic Audio Quality** of the samples and in addition you will be asked to evaluate several **sound quality attributes**. For each attribute you will also be asked to give the rating you would consider ideal for each attribute in your view. Please carefully study these attribute definitions before starting the test and ask the test administrator to clarify the terms either before the test or after the familiarization. When rating each attribute, please use the scale as you consider appropriate - you are encouraged to use the whole scale as and when needed.

At the right-hand side of each trial page, you will find an additional rating scale with an * below it. This scale is for your **ideal rating for each attribute**. Having listened to all the samples for the trial, you are asked to consider what would be the desired level for this attribute, for a hypothetical system, and show this level, by placing the slider on the scale. Please consider carefully what you have heard and your expectations when you make your **ideal rating**.

Please take your time to listen carefully and consider each attribute and ratings for each sample as well as your ideal rating. Do not rush with the 1st few samples, as the test will become easier as you progress.

Good luck with the test.

The test was split into two 2-hour sessions during which assessors were encouraged to take breaks every 30 minutes or so.
5. Data analysis

5.1 High level overview

5.1.1 Overall Means and CI (BAQ only)

Figure 3: Overall “Basic audio quality” scores per test condition, averaged over all labs, 35 assessors (post screened) and samples

![Overall “Basic audio quality” scores per test condition](image1)

Figure 4: Overall “Basic audio quality” scores per test condition, averaged over all labs, all (58) assessors and samples.

5.1.2 Raw attribute data (spider plots)

An initial perspective of the attribute can be ascertained by plotting the raw attribute data per system, averaged over all samples and assessors. Figure 5 and Figure 6 provide an overview of this data, also with the ideal profile plotted for reference. In all cases the coloured circles indicate the 95% confidence intervals for the average attribute ratings.
5.1.3 Multivariate analysis

To obtain an overview of the attribute data, principal component analysis (PCA) is applied to the data to reveal the nature of the underlying dimension of the data. Figure 7 and Figure 9 presents the system factor map which illustrates the relationship between the test conditions for principal components (PC) 1 & 2 and 2 & 3 respectively. The variables factor maps are presented in Figure 8 and Figure 10 and present the attributes which load each of the principal components. The meaning of the principle components can be understood by using the vectors of the variable factor with the system factor maps.
Figure 7: Principal component analysis (PCA), test conditions factor map for dimensions 1 and 2.

Figure 8: Principal component analysis (PCA), variables factor map for dimensions 1 and 2.

Figure 9: Principal component analysis (PCA), test conditions factor map for dimensions 2 and 3.

Figure 10: Principal component analysis (PCA), variables factor map for dimensions 2 and 3.
5.2 Detailed analysis

5.2.1 Average test condition performance per attribute

Figure 11: “Clarity” scores for per conditions, averaged over all samples, labs assessors.

Figure 12: “Envelopment” scores per test conditions, averaged over all samples, labs assessors.

Figure 13: “Localisation accuracy” scores per test conditions, averaged over all samples, labs assessors.
5.2.2 Impact of samples on system performance

Figure 14: “Scene depth” scores per test conditions, averaged over all samples, labs assessors.

Figure 15: “Tone colour” scores per test conditions, averaged over all samples, labs assessors.

Figure 16: “Basic audio quality” scores for samples*test conditions, averaged over all labs and assessors

Figure 17: “Clarity” scores for samples*test conditions, averaged over all labs and assessors
5.3 Data quality

This section reviews the data quality from a number of different perspectives, including assessor performance (§ 5.3.1), and repeatability of the data collection within each laboratory (§ 5.3.2) and overall per test condition (§ 5.3.3).

5.3.1 Assessor screening

In order to study the performance of assessors over replicated ratings of the test conditions for 2 samples, the eGauge method was applied, as reported in ITU-R Report BS-2300 [7]. The method statistically evaluates the degree of discrimination and reliability that assessors have in making their ratings. A 95% non-parametric permutation test is applied to establish a threshold of acceptance for assessors both for discrimination and reliability scores. These thresholds are indicated by the grey crosshairs in the figure below. Assessors lying in the top-right hand quadrant are those who are able to discriminate the stimuli reliably. All other assessors fall into the bottom-right hand quadrant, indicating the where reliable, but not able to discriminate the differences well, i.e. with more than a >5% error rate.

Using this metric, 34 of the 58 assessor were included in all analysis presented in this report, except for § 5.1.1, where the data from 58 assessors² is compared to that of 34 assessors in Figure 3 and Figure 4 to illustrate the similarity of the data.

² Note that assessor labels run from EBU1 - EBU83 for anonymity.
5.3.2 Repeatability within lab

To study further the repeatability of the results within each lab, the replicated sample data is analysed per laboratory, as illustrated in Figure 23 - Figure 28. A measure that the data is repeatable within each laboratory is to study whether the 1st (blue bar) and 2nd (orange bar) replication confidence intervals overlap for each in each laboratory. This is the case for all attributes and all laboratories, whilst overall difference between laboratories can be observed.
5.3.3 Repeatability per test condition

The two replicated samples were evaluated for all test conditions in all labs. One measure of the stability and repeatability of the data is to look at whether there is any significant difference between replicate 1 and 2 for each test condition and for each attribute. This is illustrated in Figure 29 - Figure 34, by comparing for each test condition the means and confidence interval overlap for the 1st (blue bar) and 2nd presentation (orange bar). For all attribute, all test conditions and the ideal ratings, there is no significant difference between the replicate for each test condition indicating that a high degree of replicability of results.
Figure 29: Average “Basic audio quality” scores for test conditions, for the 1st and 2nd replication, averaged over all labs for the 2 replicated samples.

Figure 30: Average “Clarity” scores for test conditions, for the 1st and 2nd replication, averaged over all labs for the 2 replicated samples.

Figure 31: Average “Envelopment” scores for test conditions, for the 1st and 2nd replication, averaged over all labs for the 2 replicated samples.

Figure 32: Average “Localisation accuracy” scores for test conditions, for the 1st and 2nd replication, averaged over all labs for the 2 replicated samples.

Figure 33: Average “Scene depth” scores for test conditions, for the 1st and 2nd replication, averaged over all labs for the 2 replicated samples.

Figure 34: Average “Tone colour” scores for test conditions, for the 1st and 2nd replication, averaged over all labs for the 2 replicated sample.
6. Summary of Results

The purpose of the EBU-ADM Renderer listening test was to evaluate the performance of renderers in a range of broadcast and standard listening rooms, in different loudspeaker channel configurations. An MS-IPM experiment was designed to compare 7 test conditions with 6 programme items. The test conditions comprised of a combination of channel layouts (0+2+0, 0+5+0, 4+7+0, 9+10+3), in accordance to ITU-R BS.2051-1 [4], two different renderers and one down-mix. The identical double-blind test design was performed in 5 different laboratories, comprising of either ITU-R BS.1116-3 [5] compliant listening rooms or broadcast listening labs. The user interfaces and attributes were translated into the local languages. 6 programme items were selected to represent a broad range broadcast content including, sports, radio dramas, classical and electronic music. The MS-IPM study required assessors to evaluate the performance of each test condition for each of the programme items, using 6 rating scales: basic audio quality, scene depth, envelopment, localisation accuracy, clarity, tone colour. Additionally, for each attribute the assessors were asked to envisage the ideal performance they might desire and provide a rating of this ideal level for each attribute.

As the primary research question of the study was to investigate the similarity of renderers in a range of different and pertinent listening conditions (different listening rooms, loudspeaker types and equalisation strategies), we only studied the overall performance across all laboratories.

In total 58 assessors participated in the study across the 5 laboratories. A strict post-screening was performed for basic audio quality rating using the method provided in Report ITU-R BS.2300 [7] and the best 34 assessors were included for the subsequent analysis.

The analysis was applied to study the research question for basic audio quality and each attribute individually. Additionally, a combined analysis was conducted using a multivariate analysis (PCA). When averaging across all assessors, laboratories and programme items, statistically significant differences are shown between the 7 test conditions. For basic audio quality, scene depth, envelopment, localisation accuracy, and clarity the test condition differences are primarily between the channel layouts. For any given channel layout or attribute, no statistically significant differences are found between renderers. However, with the exception of tone colour, all attributes are clearly and reliably discriminating system differences.

The multivariate analysis provides an overview of dominating perceptual characteristics and differences in the dataset. The first dimension dominates 55% of the variance of the data, whilst the second dimension shows 20% of the variance linked to tone colour. The 95% confidence ellipses allow us to study the statistical similarity of test conditions under study. For dimensions 1 and 2, for any given channel layout or attribute, no statistically significant differences are found between renderers. The loudspeaker layouts 0+2+0 and 0+5+0 are distinguished from others in the first dimension, with 0+2+0 well separated to the negative side of the axis. Layout 9+10+3 is not well separated from the 4+7+0 layout. Furthermore, all of the test conditions are shown to be statistically significantly different from the ideal rating for dimensions 1 and 2, whilst no system was rated as excellent in terms of basic audio quality. The 3rd dimension, explains a further 11% of the variance, which differentiated the test conditions in terms of localisation accuracy and clarify versus scene depth and envelopment, with layout 0+2+0 separated from all others.

Further, interpretation of the test condition performance per programme item, averaged over all assessors and all labs, provides insight into test condition differences for each programme item. For certain, attributes, samples and speaker configurations small but significant differences can be identified between renderers. For some specific audio items, the renderers were often rated equivalent to the attribute ideal point for the 4+7+0 and 9+10+3 loudspeaker configurations, whilst for other items all conditions were often rated below the ideal.
7 Conclusions

The EBU ADM Renderer and another commercially available renderer were evaluated using the multiple stimulus ideal profile method (MS-IPM) with a range of attributes, object-based audio items and loudspeaker layouts in a large-scale listening experiment across multiple laboratories. No reference signals were presented, so system ratings are with reference to the assessors’ expectation of performance, and they also reported the ideal point on attribute scales.

The results show that differences between renderers are not significant for any loudspeaker layout when averaged across programme items. Further a static downmix from loudspeaker layout 9+10+3 to 4+7+0 was also not distinguished from either object-based rendering for 4+7+0. Some significant differences between renderers were observed on a few combinations of specific programme items, loudspeaker layouts and attributes, but these were only small. When averaged across programme items, no system was rated as excellent in terms of basic audio quality. For all attributes except tone colour, test conditions were rated significantly below the ideal point when averaged across audio items, but for some specific items the renderers were equivalent to the ideal when using three-dimensional loudspeaker layouts.

Observations on the MS-IPM evaluation method are given in Annex 1.

8 Acknowledgements

The EBU gratefully acknowledges the resources, hard work and intellectual effort put into this test by the four EBU Member R&D laboratories (BBC, RAI, IRT and FT) and by FORCE Technology, SenseLab of Denmark and NHK of Japan. Without their generosity, this work would not have been possible. This work was co-funded by The Danish Council for Technology and Innovation.

9 References

[1] EBU Tech 3388 - the EBU ADM Renderer (EAR)  
https://tech.ebu.ch/publications/adm-renderer-for-use-in-nga-broadcasting

https://www.itu.int/rec/R-REC-BS.2076/en

[3] EBU Tech 3387 - MS IPM - Multiple Stimulus Ideal Profile Method  
https://tech.ebu.ch/publications/msipm-subjective-audio-testing-methodology (temp link)

https://www.itu.int/rec/R-REC-BS.2051/en

https://www.itu.int/rec/R-REC-BS.1116/en

https://www.itu.int/rec/R-REC-BS.1534/en


[16] N. Zacharov, C. P. Volk and T. S. Andersen, Comparison of hedonic and quality rating scales for perceptual evaluation of high- and intermediate- quality stimuli, in 143rd AES Convention, 2017


[19] EBU R128 - Loudness normalisation and permitted maximum level of audio signals, 2014


Annex 1: Observations for improvement of the MS-IPM

The multiple stimulus ideal profile method (MS-IPM) allowed for detailed evaluation of multiple object-based audio renderers across a wide range of conditions and attributes, when no defined reference behaviour was available. Standardised methods for subjective audio evaluation are not well suited to this task and the MS-IPM provided valuable insights into system characteristics. From the experience of this study some recommendations are made for future uses of the method.

*Experimental design:* Several assessors found it tedious to listen to the same sample for several trials in a row. Randomisation of the presentation order of audio items and attributes should be considered in the experimental design to avoid tedium.

*Audio items:* Some assessors commented that the audio items were sometimes too long and the scenes too complex in nature. Short samples should be employed to ensure consistency across the content.

*System loudness alignment:* The samples were loudness aligned, but some loudness differences still remained. Loudness alignment between stimuli must be performed. ITU-R BS.1770 may be employed, but the resulting loudness alignment should be verified, through listening, by the experimenter. Employing short samples may facilitate loudness alignment.

*Basic audio quality* is a global rating attribute strongly linked to assessment of differences from a reference (as found in Rec. BS.1534 and Rec. BS.1116). It would be wise to avoid usage of the same attribute for methods without a reference. The 100-point ITU continuous quality scale, with five adjectives ranging from bad-excellent, could still be used, but with the attribute “overall audio quality”, defined as: the overall impression of sound quality, encompassing all aspects of the entire sound sample.

*Training and familiarisation:* Some assessors commented that the training and familiarisation was very useful. Good written and verbal instructions for the experiment are essential, with clear and well understood attribute names and definitions. Assessors should be given sufficient time for training and familiarisation with all the test stimuli and attributes, prior to the main test, with guidelines regarding correct usage of attributes and associated scale usage.

*Attribute evaluation:* Some assessors commented on the difficulty of using certain attributes, others commented on the similarity of some attributes. Pertinent attributes must be selected to allow assessors to discriminate well the stimuli. A clear procedure for this should be defined. Ideally, the attribute selection and definitions are tested in a pilot evaluation with a limited number of expert assessors. To avoid assessor fatigue, the duration of the test and training session should ideally not exceed 4 hours in total, split into 2 - 3 test sessions, allowing assessors to take breaks when needed. The duration of an experiment should be estimated in advance. The required number of attributes for any given experiments should be established for the specific experimental applications, in the range of 3 - 8. Experience shows that more than 8 attributes in listening tests does not yield improved data quality.

*Ideal profile:* Some assessors found it challenging to evaluate the ideal and requested for guidance on the usage of the ideal rating scale. The confidence intervals for the ideal ratings were similar to those for the physical systems for each attribute however. Clear definition of the meaning and usage of the ideal is needed, including sufficient time for explanation and training. Assessors should be familiar with the field of application of systems under test, such that their expectations are based on experience. The ideal rating might be an optional part of a recommended method.

*Assessor performance:* Post-screening showed that nearly all assessors provided reliable ratings, but some were not discriminating on all attributes. Expert assessors with listening experience of the technology under evaluation should be employed in such tests and sufficient time for training and familiarisation should be given. In this experiment 35 post-screened assessors yielded stable and interpretable results. Analysis performed with the best performing 21 assessors yielded similar inferences.