EBU – TECH 3339

EBU Evaluations of Multichannel Audio Codecs

Phase 3

Source: D/MAE

Geneva
March 2010

[*] Page intentionally left blank. This document is paginated for two sided printing

Contents

[*] Page intentionally left blank. This document is paginated for two sided printing

EBU Evaluations of Multichannel Audio Codecs

Phase 3

EBU Committee	*First Issued*	*Revised*	*Re-issued*
DMC	2010		

Keywords: Multichannel Audio, Cascaded codecs

1. Introduction

There is now an increasing demand for surround sound material in the home, with many households owning 5.1 surround sound systems.

While most of the 5.1 content is currently played back from recorded media such as DVDs, there is more demand for broadcast, downloaded and streamed material. Therefore there are now bandwidth restrictions to deal with and the resulting need for multichannel audio codecs at lower bit-rates.

The EBU D/MAE (Multichannel Audio Evaluation) group has been assessing various multichannel audio codecs over the past four years. The best way of assessing audio codec sound quality is by subjective listening tests, and the group has been performing such tests using the combined effort of several broadcasting research labs.

The tests have been split into three Phases. The first two Phases covered emission codecs, and the Phase 3 tests cover cascaded codecs. The report for Phases 1 and 2 is included in EBU Tech 3324 [1]. The present document is the report of the Phase 3 tests comparing cascaded audio codecs.

Chains of higher bit-rate distribution or contribution codecs concatenated with an emission codec were tested. The distribution codecs used were Dolby E, E-aptX and Linear Acoustic Stream Stacker. The emission codecs used were HE-AAC, Dolby Digital and Dolby Digital Plus, with DTS and Dolby Digital employed as transcoding codecs. One of the aims was to see if the presence of several distribution codecs would have an adverse effect on the coding quality of an emission codec.

Whereas in Phases 1 and 2 the MUSHRA test method [3] was used, the BS.1116 [2] test method was used in Phase 3. The main reason for this change of method was because it was felt that the quality being assessed would be high and the differences between cascades would be small. The BS.1116 subjective test method can provide very precise results at the high quality end of audio coding albeit that it is more time-consuming to perform than MUSHRA.

No new test material was required for the Phase 3 tests, as none of the codecs were developed after the Phase 1 and 2 tests were completed. The results showed that the emission codecs (which usually operate at a far lower bit-rate to the distribution codecs) were the main influence on quality. However, there were still significant, albeit smaller, differences between the different distribution codec cascades.

2. Participating Test Sites

As the number of measurements involved in these tests was enormous, a single laboratory would not have been able to perform all the tests required in a reasonable time scale. The work was therefore divided among the following EBU Member laboratories[1]:

- Institut für Rundfunktechnik (IRT)
- Sveriges Radio (SR)
- Westdeutscher Rundfunk (WDR)

Each laboratory performed part of the overall test workload such that each part of the tests was performed by at least two laboratories. In this way it was possible to assess whether or not the results from different laboratories were sufficiently well correlated. The ITU-R BS.1116 method allows such sharing among several laboratories.

Appendix 3 describes the listening rooms and the equipment used at the laboratories.

Where a laboratory did not have enough assessors available to carry out subjective tests, the remaining sessions were carried out by another laboratory.

All laboratories were committed to carry out their tests in accordance with the relevant ITU Recommendations [2], [3] and [5].

It is important to stress the collective nature of these EBU efforts, in which solidarity and cooperation of the D/MAE members was essential to achieve the common objective of performing subjective evaluations of several multichannel audio codecs.

3. Codecs Selected for Testing

A variety of multichannel audio (MCA) codecs are currently used in radio, TV and the internet. It is likely that chains of different codecs (using varying parameters) will be used in the broadcast chain consisting of the studio, contribution, distribution, emission and also in the home environment. Some hypothetical multichannel audio codec chains are depicted in Figure 1.

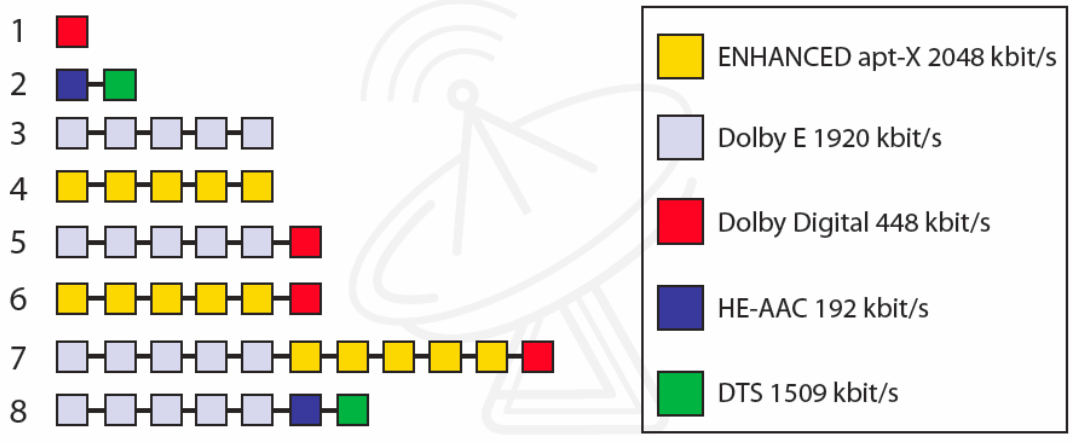


Figure 1: Some hypothetical multichannel audio codec chains in broadcasting.

[1] Two other EBU laboratories, Radiotelevisione Italiana (RAI CRIT) and Bayerischer Rundfunk (BR), were participating in the preliminary tests but their results could not be used in the final statistical analysis. Their descriptions are also included in **Appendix 3**.

The present EBU tests involved several MCA codecs fulfilling one of the two following criteria:

- they must be standardised by the DVB Consortium (TS 101 154), or
- Be commercially available and/or attractive for use in broadcasting.

In order to identify the codecs corresponding to the above conditions, D/MAE analysed the whole broadcast chain, including production, contribution, distribution and emission, as used by the participating EBU member organisations, and identified possible MCA codecs. For details, see EBU Tech 3324 [1].

The following fifteen MCA codec chains were tested in Phase 3 (Table 1):

Table 1: Phase 3 MCA codec chains tested

No	MCA codec chain
1	Original
2	Dolby Digital 448 kbit/s
3	Dolby Digital 640 kbit/s
4	Dolby Digital Plus 256 kbit/s – Dolby Digital 640 kbit/s
5	HE AAC 160 kbit/s – DTS 1509 kbit/s
6	5x Dolby E – Dolby Digital 448 kbit/s
7	5x Dolby E – DDP 256 kbit/s – Dolby Digital 640 kbit/s
8	5x Dolby E – HE AAC 160 kbit/s – DTS 1509 kbit/s
9	5x Dolby E – 5 E-aptX - Dolby Digital 448 kbit/s
10	5x E-aptX – Dolby Digital 448 kbit/s
11	5x E-aptX – DDP 256 kbit/s – Dolby Digital 640 kbit/s
12	5x E-aptX – HE AAC 160 – DTS 1509 kbit/s
13	5x Linear Acoustic - Dolby Digital 448 kbit/s
14	5x Linear Acoustic – Dolby Digital Plus 256 kbit/s – Dolby Digital 640 kbit/s
15	5x Linear Acoustic – HE AAC 160 kbit/s – DTS 1509 kbit/s

The detailed descriptions of the distribution/contribution codecs used in these tests (i.e. aptX and Dolby E) are given in **Appendix 2**. For Linear Acoustic see the note on pages 8 and 9.

4. Codec Parameters

The principal MCA codec parameters used in the tests were as follows:

Audio mode: The selection of test material included both 5.0 and 5.1 channel audio modes.

Sampling rate: 48 kHz. It was not practicable to use 96 kHz sampling rate for testing.

Input audio bit-depth: At least 16 bit, preferably 24 bit.

Encoding parameters: Pre-configured by the codec developers.

The above parameters remained consistent during the whole duration of the tests.

5. Test Sequences

The test sequences used were audio excerpts chosen from typical radio and television programme services.

Length of test sequences: Approximately 15 seconds.

Format: Unprocessed PCM (WAV or AIFF) and DSD (SACD). Dolby-E encoded and material processed using broadcast processors (such as Optimod) were also permitted.

Pre-selection process (Phases 1 and 2): In Phase 1 and 2 the subjective tests were preceded by a pre-selection process, during which the number of test sequences was reduced down to ten per Phase. In addition, the pre-selection panel selected four items to be used for training. The items selected for test and training in Phase 3 were extracted from the set of items used in Phases 1 and 2 (see Table 2). [1]

Table 2: Phase 3 Audio sequences selected for test and training sessions

No	Test items	Description	Used in
1	Applause	Applause with distinct clapping sounds	Phase 1
2	Bach_organ2	Church organ	Phase 1
3	Harpsichord	Solo harpsichord; isolated notes	Phase 1
4	Moonriver_Mancini	Mouth organ and string orchestra	Phase 1
5	R_Plant_Rock	Rock music ("Whole lotta love")	Phase 1
6	BrassEX	Exodus; orchestral; lots of brass instruments (phase 1 training item)	Phase 2
7	Fleetwd	Transient guitar and male vocalist	Phase 2
8	HornWag	Orchestral string piece by Lohengrin	Phase 2
9	TenorRP	Radio drama; clarinet, orchestra, male speaker, tenor, ambience	Phase 2
10	Trumpet	Orchestral piece with a trumpet	Phase 2
	Training items		
1	Hadouk	Eastern woodwind, strings, percussion.	Phase 1
2	Hoelsky	Chattering choral voices.	Phase 1
3	JazzApl	Jazz Burghausen; live performance with applause.	Phase 2
4	Timpani	Orchestra with trumpet, trombone, bass drum, snappers (de Falla)	Phase 2

6. Encoding Process

6.1 Codecs

In order to subjectively evaluate the performance of codecs, all test sequences need to be encoded (and decoded) using the selected codecs at the chosen bit-rates. The encoding of all the test sequences was performed by IRT. For Phase 3 no pre-selection was done. The Enhanced aptX, Linear Acoustic and Dolby Digital Plus software encoder all required six mono wav files as inputs. The DTS software encoder required three stereo wav files (i.e. L/R, C/LFE and Ls/Rs) as inputs. These encoders operated in non-real-time.

The hardware Dolby Digital encoder required three AES digital stereo inputs. The resulting

Dolby Digital bitstream was fed directly into the Dolby Digital decoder, which produced three AES digital stereo outputs. This encoder operated in real time.

All software codecs had been installed on a Windows XP PC with a 2.4 Hz CPU and 512 Mbyte of RAM.

Table 3 shows the version number of the various codecs used to generate the test and training sequences in Phase 3:

Table 3: Codecs used in Phase 3

Codecs – Phase 3	Version/Configuration
Dolby Digital	Encoder Model DP 569, Firmware version: v.2.0.3.1
	Decoder Model DP 562
Dolby Digital plus	Dolby Digital plus Prototype Broadcast Encoder Version 1.6.0
	Dolby Digital plus Decoder-Converter Simulation Version 1.1.7
AAC / HE-AAC	Coding Technologies aacPlus v2 Evaluation Encoder V8.0.1
	Coding Technologies aacPlus v2 Evaluation Decoder V8.0.1
Dolby E	Encoder: Model DP 571
	Decoder: Model DP 572
	Adjustment: 5.1 + 2 = 8 channels (20 bit/48 kHz AES-channel)
	Bit-rate: 1920 kbit/s
WorldNet Oslo (enhanced apt-X)	Bit-rate 5.1 + 1 stereo channels = 1.536 Mbit/s
	Bit-rate per stereo pair = 384 kbit/s
	8 channels 384 kbit/s *4 = 1.536 Mbit/s
	8 channels ,16 bit, 22 kHz -> 24 time slots at 64 kbit/s
Linear Acoustic Stream Stacker Unit (see the Note below)	Adjustment: 5.1 + 2 + 5.1 + 2 =16 channels
	(1920 kbit/s = 20 bit/48 kHz AES-channel)
	Bit-rate: 960 kbit/s related to 8 channels
DTS	DTS HD Pro Series Encoder v0.97
	DTS HD Decoder Library Version 300.38

NOTE about Linear Acoustic Stream Stacker:

> *The Linear Acoustic Stream Stacker unit used in these tests has been provided by the J+C Intersonic AG, Munich office. The use of this Stream Stacker unit could not be endorsed by Linear Acoustic prior to the tests. To this end, the test results given in this document relating to LA Stream Stacker are only indicative and do not prejudice the current performance quality level of the Stream Stacker.*
>
> *Following the EBU tests, Linear Acoustic have switched to the apt-x codec to improve error resiliency, lower latency and provide up to 16 full bandwidth phase locked audio channels, something the previous codec was not capable of. The implementation remains the same with hardware encoders and decoders, OEM encode/decode modules, and licensed software decoding. Linear Acoustic now refers to the system by the name of the format: "e-squared".*

In the remainder of the report the codec names will be abbreviated in graphs and tables. Here are the abbreviations used (see Table 4):

Table 4: Notations used

Abbreviation	Full name
DD	Dolby Digital
DD+ or DDP	Dolby Digital Plus
HE-AAC	High Efficiency – Advanced Audio Codec
E-aptX	WorldNet Oslo (enhanced apt-X)
Lin	Linear Acoustic Stream Stacker Unit
DE	Dolby E

6.2 Verification of bit-rates

To verify how close the actual operating bit-rates of the encoders were to their selected bit-rates, the encoded bit-streams were analysed. Verification was performed for each software codec, at each of its bit-rates and for each test item.

In Phase 3, bit-rates were calculated using the following formula:

*Coded file/kbyte * 8 / length of the sequence/s = bit-rate kbit/s*

Verification of bit-rates in Phases 2 and 3 were performed in a slightly different way. Instead of checking the bit-rate of each individual test and training item separately, the bit-rates of all test and training sequences were verified in a batch, which made the whole verification process much faster.

The following table (Table 5) shows the difference between the selected and measured bit-rates for each codec under test.

Table 5: Target and measured bitrates

Codec chain under test	Target bitrate [kbit/s]	Measured bitrate [kbit/s]
Dolby Digital 448 kbit/s	448	448
Dolby Digital 640 kbit/s	640	640
Dolby Digital Plus 256 kbit/s – Dolby Digital 640 kbit/s	640	640
HE AAC 160 kbit/s – DTS 1509 kbit/s	1509	1501
5 Dolby E – Dolby Digital 448 kbit/s	448	448
5 Dolby E – DDP 256 kbit/s – Dolby Digital 640 kbit/s	640	640
5 Dolby E – HE AAC 160 kbit/s – DTS 1509 kbit/s	1509	1509
5 Dolby E – 5 E-aptX - Dolby Digital 448 kbit/s	448	448
5 E-aptX – Dolby Digital 448 kbit/s	448	448
5 E-aptX – DDP 256 kbit/s – Dolby Digital 640 kbit/s	640	640
5 E-aptX – HE AAC 160 – DTS 1509 kbit/s	1509	1512
5 Lin - Dolby Digital 448 kbit/s	448	448
5 Lin – Dolby Digital Plus 256 kbit/s – Dolby Digital 640 kbit/s	640	640
5 Lin – HE AAC 160 kbit/s – DTS 1509 kbit/s	1509	1509

7. Experimental design

7.1 Test method – Phases 1 and 2

The test method adopted by D/MAE for Phases 1 and 2 was MUSHRA ("MUlti Stimulus test with Hidden Reference and Anchors") [3]. This method provides a successful approach for assessing and grading different impairments. The scale for the grading was based on a video signal evaluation method (ITU-R BT.500), where the intervals are labelled, "bad", "poor", "fair", "good" and "excellent". The value at the lower end of the scale is zero; the value at the upper end is 100. The method uses the unprocessed original test item with full bandwidth as the reference signal. The set of stimuli to be assessed consists of a hidden reference, at least two anchor signals, and the signals processed by the codecs under test.

D/MAE made a decision to use the MUSHRA methodology for all the tests in Phases 1 and 2 because it covers the whole quality range and is less time-consuming than ITU-R BS.1116 testing.

7.2 Test method – Phase 3

For Phase 3, D/MAE decided to use ITU-R BS.1116 ("Methods for the Subjective Assessment of Small Impairments in Audio including Multichannel Sound Systems") [2]. ITU-R BS.1116, also known as "double-blind triple-stimulus with hidden reference", has been found to be especially sensitive, stable and to permit accurate detection of small impairments. It was expected that the codecs under test in Phase 3 mostly generate such small impairments; therefore ITU-R BS.1116 was considered to be the appropriate method.

In the preferred and most sensitive form of this method, one subject at a time is involved, and the selection of one of three stimuli ("REF", "A" and "B") is at the discretion of this subject. The known reference is always available as stimulus "REF". The hidden reference and the test stimuli are simultaneously available but are randomly assigned to "A" and "BC" for each trial.

The subject is asked to assess the impairments on "A" compared to "REF", and "B" compared to "REF", according to the continuous five-grade impairment scale. One of the stimuli, "A" or "B", should be indistinguishable from stimulus "REF"; the other one may reveal impairments. Any perceived differences between the reference and the other stimuli must be interpreted as impairments.

The excerpt may be listened to repeatedly until the subject has made an assessment they are content with. As soon as the subject has completed the grading of a trial, it should be possible to proceed directly on to the next trial. In this way the test procedure is self pacing.

The grading scale shall be treated as continuous with "anchors" derived from the ITU-R five-grade impairment scale given in Recommendation ITU-R BS.1284 (Table 6).

Table 6: Comparison of Impairment scale (ITU-R BS.1116/1284) and Quality scale (ITU-R BS.1534/1284)

Impairment scale ITU-R BS.1116/1284	Grade	"diff-grade" (Reference = 5.0)
Imperceptible	5.0	0.0
Perceptible, but not annoying	4.0	-1.0
Slightly annoying	3.0	-2.0
Annoying	2.0	-3.0
Very annoying	1.0	-4.0

Quality scale ITU-R BS.1534/1284	**Grade range ITU-R BS.1534**
Excellent	81 -100
Good	61 – 80
Fair	41 – 60
Poor	21 – 40
Bad	0 – 20

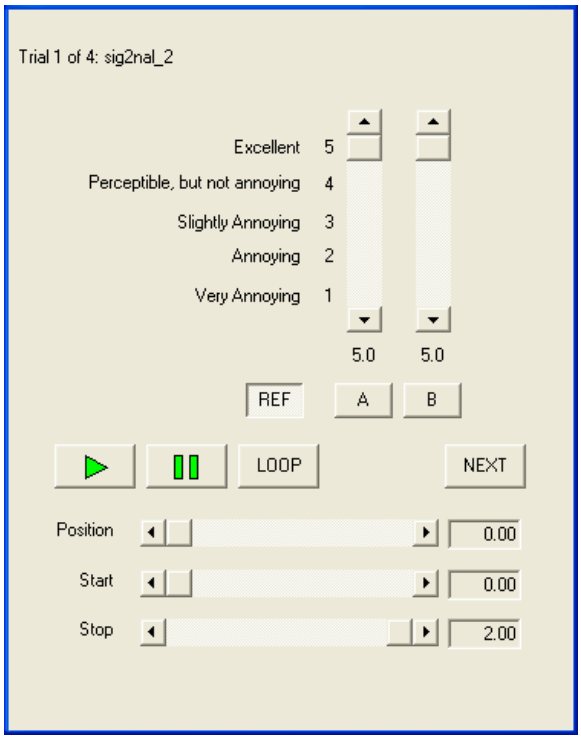


Figure 2: Graphical user interface of the ITU-R BS.1116 ARL STEP software package

Figure 2 shows the graphical user interface of the ITU-R BS.1116 ARL STEP software package (Version 1.05) ("Subjective Training and Evaluation Program (STEP) - A computer-controlled system for audio presentation and subjective evaluation" by Audio Research Labs).

7.3 Impairment (artefact) categories for MCA tests

For each coded sequence the subjects should produce a single aggregate assessment. Such an assessment embraces many parameters (see Table 7 below). The subject should decide how to weight these parameters according to their preference.

Table 7: Impairment categories

No.	Artefact category	Explanation
1	Signal correlated noise	Coloured noise associated with the signal
2	Loss of high frequency	Lack of high frequencies, dull sounding
3	Excess of high frequency	Excess of high frequencies or associated effects, e.g. sibilance or brightness
4	Periodic modulation effects	Periodic variations such as warbling, pumping, or twitter
5	Temporal distortion	Pre- and post-echoes, smearing, effects associated with transients
6	Distortion	Harmonic or inharmonic distortion
7	Loss of low frequency	Lack of low frequencies
8	Image quality	All aspects including narrowing, spreading, movement and stability
9	High frequency distortion	Distortion in the high frequencies, including phase distortions

7.4 Evaluation Process

The first step in the listening test is the familiarization with the listening test process. This phase is called a training phase and it precedes the true evaluation phase. The purpose of the training phase is to allow to the subject to achieve two objectives as follows:

- PART A: to become familiar with all the multichannel audio items, both the reference and coded versions. The listening level could be adjusted at this stage to a comfortable setting.
- PART B: to learn how to use the test equipment and the grading scale by means of 4 specially selected multichannel audio training items (not to be included in the main test).

In PART B of the training phase the subject was able to listen to all 4 multichannel audio training items at the different possible degradations in order to illustrate the whole range of possible qualities. Similar to the test items, these training items were more or less critical depending on the bit-rate and other conditions used (such as the codecs used). In this part of the training phase the subject was asked to use the available scoring equipment and to evaluate the impairment of the items by inputting the appropriate scores on the impairment scale.

Test instructions were handed out to the prospective assessors before they started to carry out the subjective evaluations. It is important that the test instructions are agreed by all participating laboratories, so that they are implemented in the same way. In order that the test instructions are correctly understood by all assessors, it may be useful to present them with a copy in their local language (e.g. Italian, French).

The purpose of the grading phase is to score the items across the given impairment scale. The subject scores should reflect their subjective judgment of the impairment level for each multichannel audio item that is presented. Each item is approximately 15 seconds long.

The subjects should listen to the reference and the test conditions by clicking on the respective buttons. They are allowed to listen to the signals in any order, any number of times. Repeated playback is available thanks to a “Loop” button. If the subject wants to focus on a certain part of the multichannel audio item, they are allowed to select this by changing the start and end markers. The subject is allowed to adjust these markers only after listening to the whole multichannel audio item. A slider for each signal is used to indicate the subject’s opinion of the current signal impairment. Once the subject is satisfied with their grading, they click on the “Trials” or “Next” button at the bottom of the screen in order to get the next trial.

The impairment scale, shown in Table 6 was used. The scale has to be interpreted as continuous from "*Very annoying*" (1.0) to "*Imperceptible*" (5.0).

When evaluating the items, the subject does not necessarily give the grade "*Very annoying*" to the item with the lowest quality in the test. However, one or more items must be given the maximum grade "*Imperceptible*" because the unprocessed reference signal is included as one of the two multi-channel audio items to be graded. During the training phase, the subjects should be able to learn how to interpret the audible impairments in terms of the grading scale. They may be encouraged to discuss personal interpretation with the other subjects at any time during the training phase. No grades given during the training phase will be taken into account in the true tests.

7.5 Listening conditions

In order to obtain comparable results that can be used by the same statistical model, the listening conditions of all the participating laboratories should be aligned in terms of equipment used as much as possible.

Appendix 3 describes the listening conditions and audio test systems used in the different EBU laboratories.

7.6 Test Sessions

The listening tests had to provide sufficient statistical coverage of each of the codecs under test, aiming for at least 15 listeners per codec. The number of codecs each listener should listen to was chosen to be 4 for these BS.1116 tests. This was considered a sensible number of stimuli for each listener: not too many to cause fatigue or confusion, but enough to get sufficient coverage.

Each listener would cover all 10 test items, along with 4 training items.

To ensure that variations between labs' scoring did not cause problems, it was important that each lab tested all of the codecs. To ensure this occurred, each listener received an individual session file with a different combination of codecs.

The codecs were split into three groups. The first group contained the 4 emission-only codecs. The 10 cascades were then split into two groups of 5 to make the second and third groups. Each session contains at least one codec from each group. This ensures a wide range of quality for each listener, and hopefully some cascades that can be identified as different from the original.

Each session file was generated with a pseudo-random combination of codecs from the three groups, so that each listener received a different file. The codecs covered by all the session files were counted to ensure each codec was evaluated by at least 15 listeners.

The session files were generated for use with the ARL STEP software, which the all of the test labs were using. These files were ASCII text files containing a list of the required file combinations for each test item. The files were generated by a simple C program that gave each session file it generated a pseudo-random combination of four cascades for all ten test items. The ARL STEP software automatically randomizes the order stimuli under test and the sequence of the test items, so this did not have to be dealt with in the session files.

The 14 cascades are grouped as shown in Table 8.

Table 8: Grouping of cascades

Group 1	Group 2	Group 3
DD 448	DE x5 -> DD 448	E-aptX x5 -> DD+ 256 -> DD 640
DD 640	DE x5 -> DD+ 256 -> DD+ 640	E-aptX x5 -> HE-AAC 160 -> DTS 1500
DD+ 256 -> DD+640	DE x5 -> HE-AAC 160 -> DTS 1500	Lin x5 -> DD 448
HE-AAC 160 -> DTS 1500	DE x5 -> E-aptX x5 -> DD 448	Lin x5 -> DD+ 256 -> DD 640
	E-aptX x5 -> DD 448	Lin x5 -> HE-AAC 160 -> DTS 1500

8. Statistical Analysis and Post-Screening

The test results are presented in terms of means of diff-grades and confidence intervals for the means. The confidence intervals give a range of values around the mean where you expect the "true" (population) mean to be located with a given level of certainty. In the presented results the level of certainty has been chosen to be 95%.

The ranges defined by the confidence intervals are therefore appropriate to give an estimate of the significance of mean differences when comparing different results. In other words, with overlapping confidence intervals it can be concluded that the observed means do not differ significantly. On the other hand, with non-overlapping confidence intervals it can be concluded that the means differ significantly.

In order to guarantee reliable test results in subjective tests like these, post-screening of subjects is necessary. Two rejection criteria were devised. All the listeners' results were checked with these rejection criteria. If any of the listeners failed any of these, they were removed from the results. The two criteria were:

1) Analysing diff-grades of all codec evaluations against the individual listener's diff-grades. Test the "null-hypothesis" by means of a t-test. Rejection criterion: t-score > -2.0. This determines whether a listener significantly differs from the trend of the whole population of listeners.

2) Analysing the average diff-grades of the hidden reference for each listener. Rejection criterion: mean diff-grade < -0.25. This determines whether a listener can reliably identify the hidden reference or whether they could just be guessing.

The result of criterion 1 was that one subject had to be rejected.

The result of criterion 2 was that six subjects had to be rejected. The corresponding diff-grades for these subjects are presented in Figure 3.

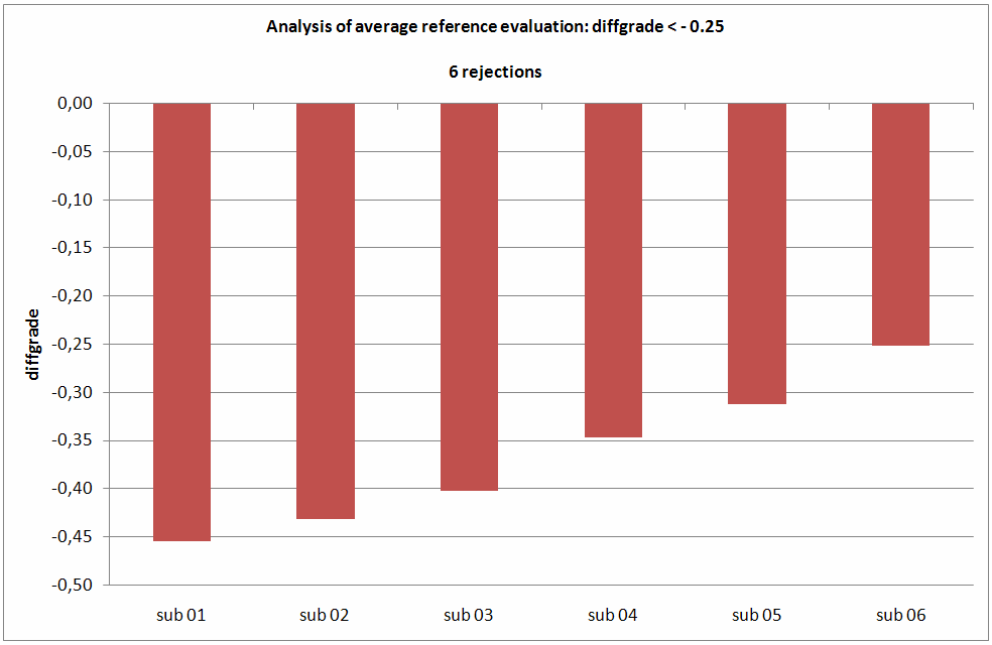


Figure 3: Analysis of average hidden reference evaluation. Diff-grades of rejected subjects

The number of number of individual listeners per codec and item, after rejection, is presented in Table 9.

Table 9: Number of individual listeners per codec and item (after rejection of subjects)

	Applause	BrassEx	fleetwd	Harpsichord	Hornwag	Mancini	organ2	Rock	tenorRP	trumpet	all items
02_DD448	21	21	21	21	21	21	21	21	21	21	210
03_DD640	19	19	19	19	19	19	19	19	19	19	190
04_DDP256 - DD640	18	18	18	18	18	18	18	18	18	18	180
05_HEAAC160 - DTS1500	17	17	17	17	17	17	17	17	17	17	170
06_DE5 - DD448	17	17	17	17	17	17	17	17	17	17	170
07_DE5 - DDP256 - DD640	19	19	19	19	19	19	19	19	19	19	190
08_DE5 - HEAAC160 - DTS1500	20	20	20	20	20	20	20	20	20	20	200
09_DE5 - aptX5 - DD448	17	17	17	17	17	17	17	17	17	18	171
10_aptX - DD448	17	17	17	17	17	17	17	17	17	17	170
11_aptX5 - DDP256 - DD640	18	18	18	18	18	18	18	18	18	18	180
12_aptX5 - HEAAC160 - DTS1500	17	17	17	17	17	17	17	17	17	17	170
13_Lin5 - DD448	18	18	18	18	18	18	18	18	18	18	180
14_Lin5 - DDP256 - DD640	15	15	15	15	15	15	15	15	15	15	150
15_Lin5 - HEAAC160 - DTS1500	17	17	17	17	17	17	17	17	17	17	170
Sum	250	250	250	250	250	250	250	250	250	251	

9. Presentation of Main Results

The overall results for Phase 3 are presented as follows:

- average score for each codec over all test items
- average score for each item over all codecs

The detailed results for each codec tested for each and every test item are given in **Appendix 1**.

9.1 All Codecs averaged over all test items plus average of worst case item

The blue bars represent the mean diff-grades over all items. The vertical black lines, overlaying the mean diff-grades, are the 95% confidence intervals of the mean diff-grades. The white horizontal lines represent the mean diff-grade of the worst case item. The name of the worst case item is given above the codec name. Figure 4 shows the codecs in numerical order, and Figure 5 shows the codecs sorted by mean diff-grades.

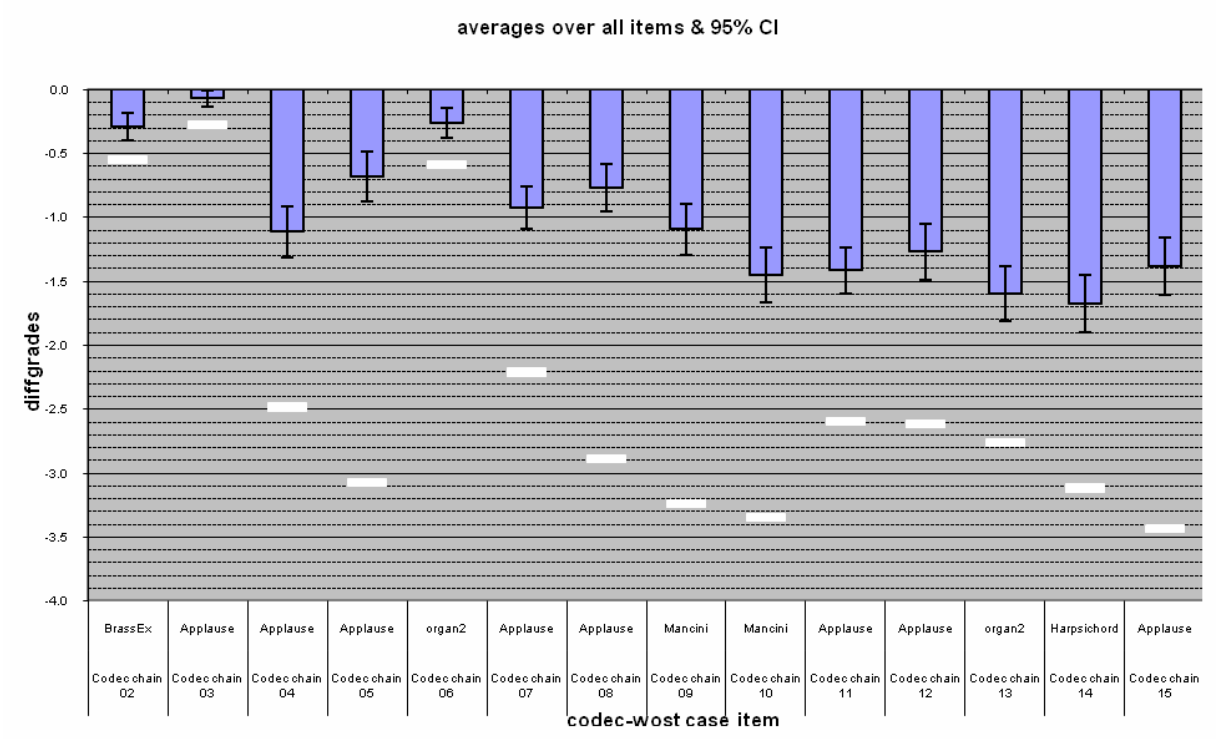


Figure 4: ITU-R BS.1116 diff-grades of all codecs averaged over all test sequences including worst case item and its diff-grade

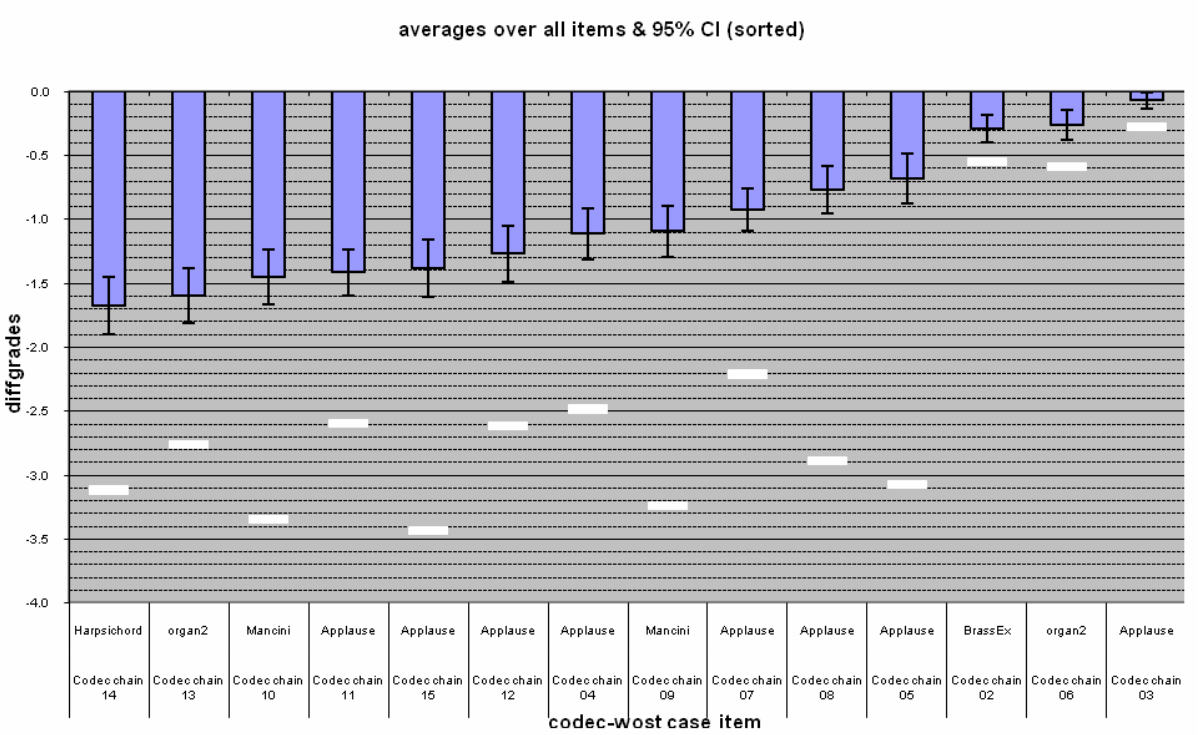


Figure 5: ITU-R BS.1116 diff-grades of all codecs averaged over all test sequences including worst case item and its diff-grade (sorted by mean diff-grade)

9.2 Average scores for items over all codecs

The graph in Figure 6 shows how critical each test item was. The mean for each item over all the codecs is shown.

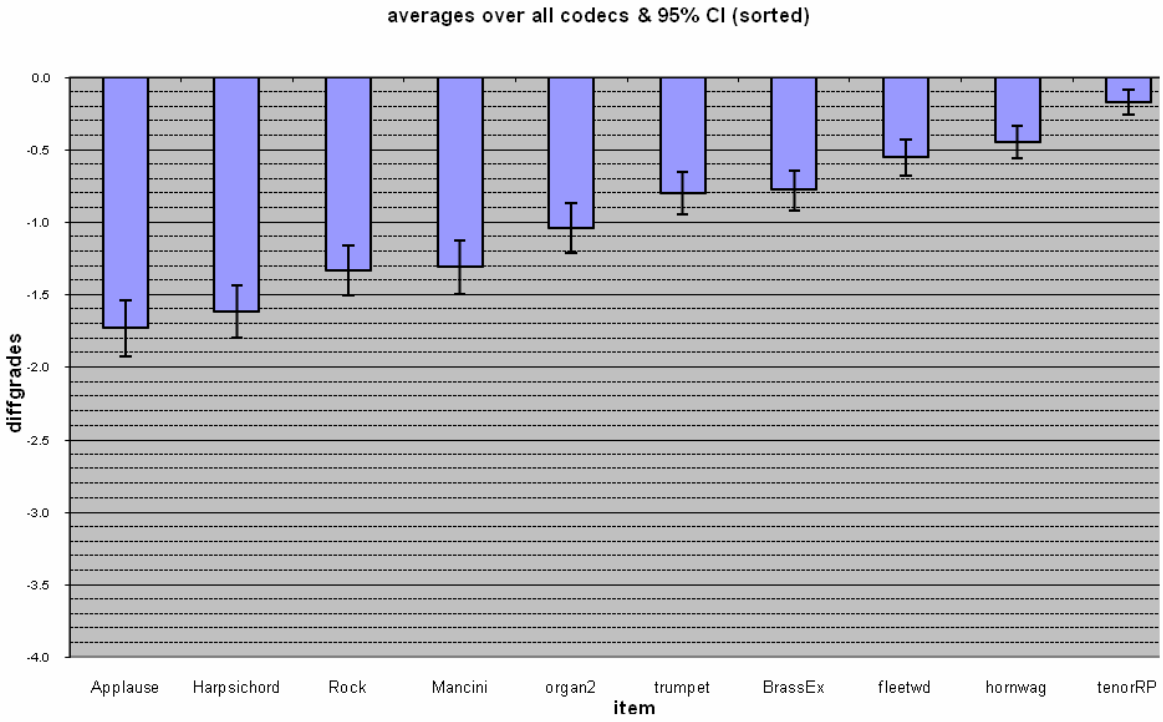


Figure 6: ITU-R BS.1116 diff-grades of all items averaged over all codecs (sorted)

9.3 Average scores for each lab over all items for four individual codecs

The graph in Figure 7 gives the average scores over all items and the 95% confidence intervals for each participating lab for four different codecs/cascades: DD448 (number 2), DD+256-DD640 (4), DE5-aptX5-DD448 (9) and Lin5-DD448 (13).

Interpreting the overlapping of confidence intervals as non-significant differences of the mean grades, the results suggest that none of the laboratories differ significantly from the overall mean grading (averaged over all labs), shown with the red bar below (marked ALL).

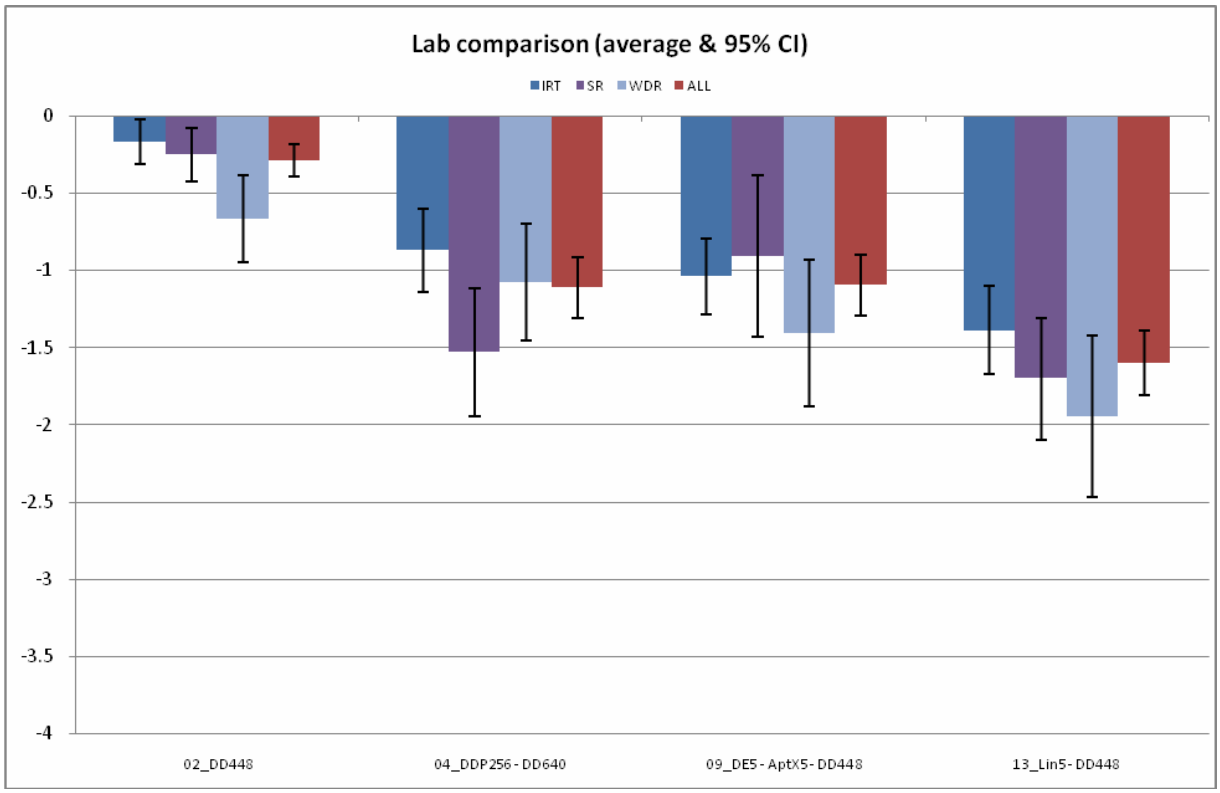


Figure 7: Comparison of the three laboratories with the overall average results

10. Summary and Conclusions

Overall, the EBU multichannel audio evaluation work highlighted some issues found with large-scale subjective tests. The test method used aimed at minimizing any differences that there may have been between test labs that may have influenced the results, and allowing analysis of these differences.

All the labs had listening rooms that conformed to the required standard, with correctly arranged loudspeakers and sufficiently high quality equipment to ensure that the physical and technical differences between the labs were insignificant. Any difference between labs is then primarily down to the individual listeners. For example, one lab may have some expert listeners who object to high frequency artefacts, whereas a second lab may have some experts who dislike low frequency distortions. Both groups of listeners may give very different but nonetheless valid scores for particular codecs. This can result in different labs giving apparently different average scores. This can be an advantage as the diversity of listeners required for good results is increased if several test labs are used.

Analysis of the results is not straightforward either. As just mentioned, each individual, even after training, practice and with good hearing, is different. Therefore perfectly valid results from two

different people may differ greatly; so it is important to have a statistically significant number of listeners for each codec under test. According to BS.1116, this number should be at least 15 valid assessors.

It is important to look not just at the average scores, but to examine other details too. For example, a codec's average score over all test items may be high (such as a -1.2 diff grade), but for one test item it may struggle (such as a -3.7 diff grade). This information is vital, particularly if that test item is something that could be typically broadcast and thus must not be ignored.

It must be remembered that the contribution codecs in these tests have been cascaded five times. In radio broadcasting this figure is typically lower, whereas in television broadcasting it can be higher. Fewer cascaded codecs should generate smaller audio impairments than the results here show, but more cascaded codecs are likely to increase impairments.

For a coded test item to be considered of a quality that is safe for broadcast use it must have a diff-grade better than -1.0. Table 10 shows the number of test items that fail this criterion for each chain. Three chains (A, B and E) do not have any test items that fail this criterion so no test item scores worse than -1.0. With the remaining chains there are test items which generate diff-grades worse than -1.0, and the graphs in **Appendix 1** show which test items produce these failures for each of the codec chains.

If cascaded codec chains are to be considered for broadcast use then the quality criterion should be that none of material should produce an average diff-grade worse than -1.0 ("*perceptible but not annoying*"). If all the tested items score better than -1.0, then we can consider the chain's performance to be acceptable. However, if the average over all items is better than –1.0, but some test items score worse than this, then we must be wary of using such a chain. If the average over all items is worse than -1.0 then the chain should not be considered acceptable for broadcast use.

The acceptability of different chains used in the tests is shown in Table 10.

Table 10: Acceptability of codec chains for broadcast use

Label	Contribution/Distribution chain	Emission codec	Number of items worse than -1.0 (out of 10)
2	Direct	DD 640	0
3	Direct	DD 448	0
4	Direct	DD+ 256 -> DD640	7
5	Direct	HE-AAC 160 -> DTS 1500	3
6	5x Dolby E	DD 448	0
7	5x Dolby E	DD+ 256 -> DD640	3
8	5x Dolby E	HE-AAC 160 -> DTS 1500	2
9	5x Dolby E -> 5x E-aptX	DD 448	3
10	5x E-aptX	DD 448	8
11	5x E-aptX	DD+ 256 -> DD640	9
12	5x E-aptX	HE-AAC 160 -> DTS 1500	6
13	5x Linear Acoustic	DD 448	7
14	5x Linear Acoustic	DD+ 256 -> DD640	8
15	5x Linear Acoustic	HE-AAC 160 -> DTS 1500	7

It must be made clear that the cascaded contribution codecs operate at different bit-rates, which are shown in Table 11. Therefore, as would be expected, the lower bit-rate codecs are producing the larger impairments. The channel configuration for the Linear Acoustic codec could have been

set to 8 channels; however, the advantage of this codec had particularly been seen in the 16 channel mode, which would allow not only 5.1 and 2.0 audio for international distribution, but extra audio channels for national distribution as well.

Table 11: Bit-rates of cascades

Contribution Codec	Bit-rate per channel [kbit/s]
Dolby E	240
E-aptX	192
Linear Acoustic	120

The results also highlight an interesting positive effect of cascading codecs. Codec chain 8 (with 5x Dolby E) out-performs chain 5 (no contribution codecs), which both have the emission codec HE-AAC. Also, chain 7 appears to be better than 4 and chain 9 beats 10. All these chains have Dolby E appearing to help improve the performance of the emission codec. A possible reason for this phenomenon is that Dolby E band-limits the very top-end of the frequency range (often beyond the range of the hearing of most listeners with most test items). These high frequencies can be problematic for the emission codecs to encode, so their removal can reduce some of the artefacts, hence the improved scores. Therefore Dolby E appears to act as a pre-processor for the emission codecs.

The results show that Dolby E operating at 240 kbit/s per channel does not degrade the performance when feeding any of the emission codecs tested. E-aptX and the Linear Acoustic codecs, which operate at lower bit-rates, do cause significant impairment and therefore ought to be avoided unless higher bit-rates can be used.

The significant bottlenecks in impairments are due to the emission codecs, particularly when looking at their worst case test items. The two lowest bit-rate codecs, DD+ at 256 kbit/s and HE-AAC at 160 kbit/s, both produce significant audio impairments with the “applause” item, which is rather a common sound in broadcasting. The lowest scoring test items for the contribution codecs were not necessarily the same items as the emission codecs. For example, the “Harpsichord” item produced the lowest score with the Linear Acoustic codec, which when combined with a lower-rate emission codec produced large impairments in both the “Harpsichord” and “Applause” items, bringing down the overall average score.

Not only have these tests provided useful test results for those interested in the codecs themselves, they have also shown good practice in performing listening tests which could be of use for others in the future.

The organisation of large scale listening tests such as this, require significant efforts and resources from the EBU members. EBU members are uniquely positioned collectively to undertake such tests and we would like to encourage the members to continue contributing work in this area.

11. Acknowledgements

Members of the EBU D/MAE (Multichannel Audio Evaluations) Project Group, chaired by Gerhard Stoll (IRT) would like to gratefully acknowledge the significant efforts of the participating laboratories BR, IRT, RAI CRIT, SR and WDR as well as TVP, NRK, SVT, Radio France and the BBC. Thanks should go to Dolby, DTS, Coding Technologies, aptX (now Audemat), Linear Acoustic and Fraunhofer IIS for their support. The Group appreciates the work of Gerhard Spikofski (IRT) for his effort in providing the statistical analysis and David Marston (BBC) who was responsible for setting up the session files as well as assisting in editorial revisions.

12. References

[1] EBU Evaluations of Multichannel Audio Codecs, EBU – Tech 3324, Geneva, September 2007

[2] ITU, "Recommendation ITU-R BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems" October 1997

[3] ITU, "Recommendation ITU-R BS.1534-1 "Method for the subjective assessment of intermediate quality levels of coding systems", January 2003

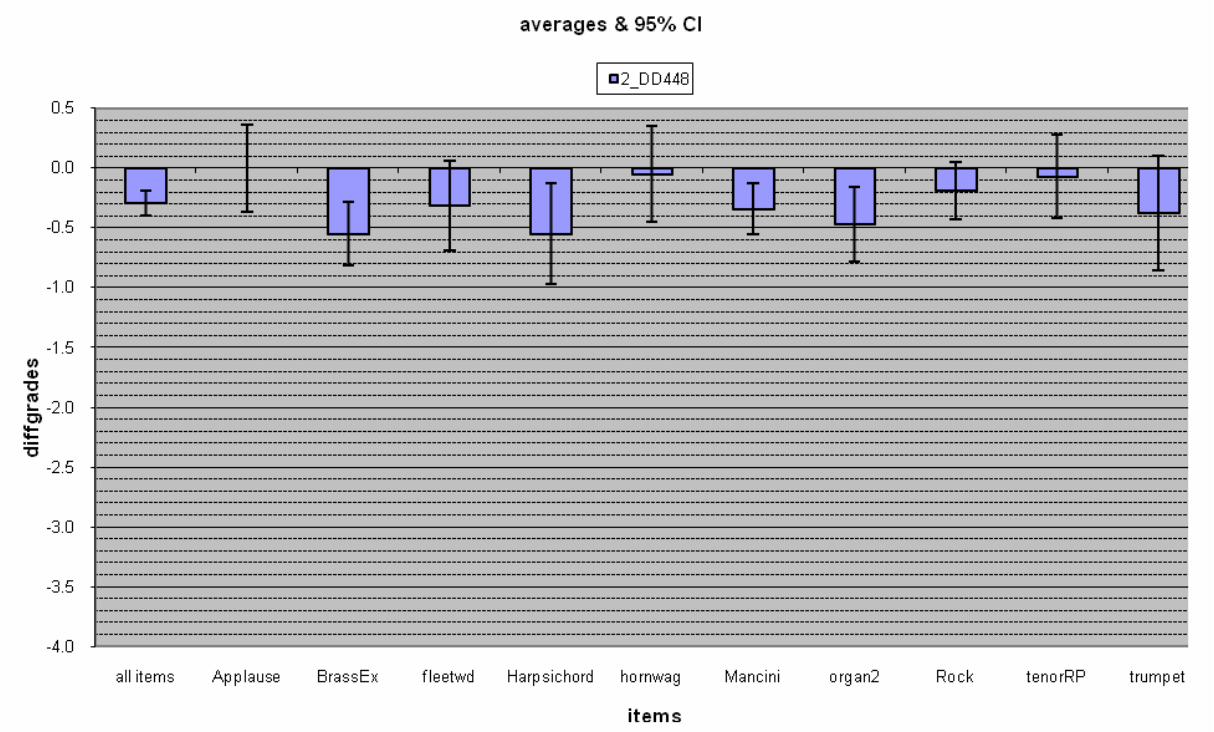
[4] EBU Tech.doc. 3309: Evaluations of Cascaded Audio Codecs, Project Group B/AIM (Audio in Multimedia), Geneva, June 2005

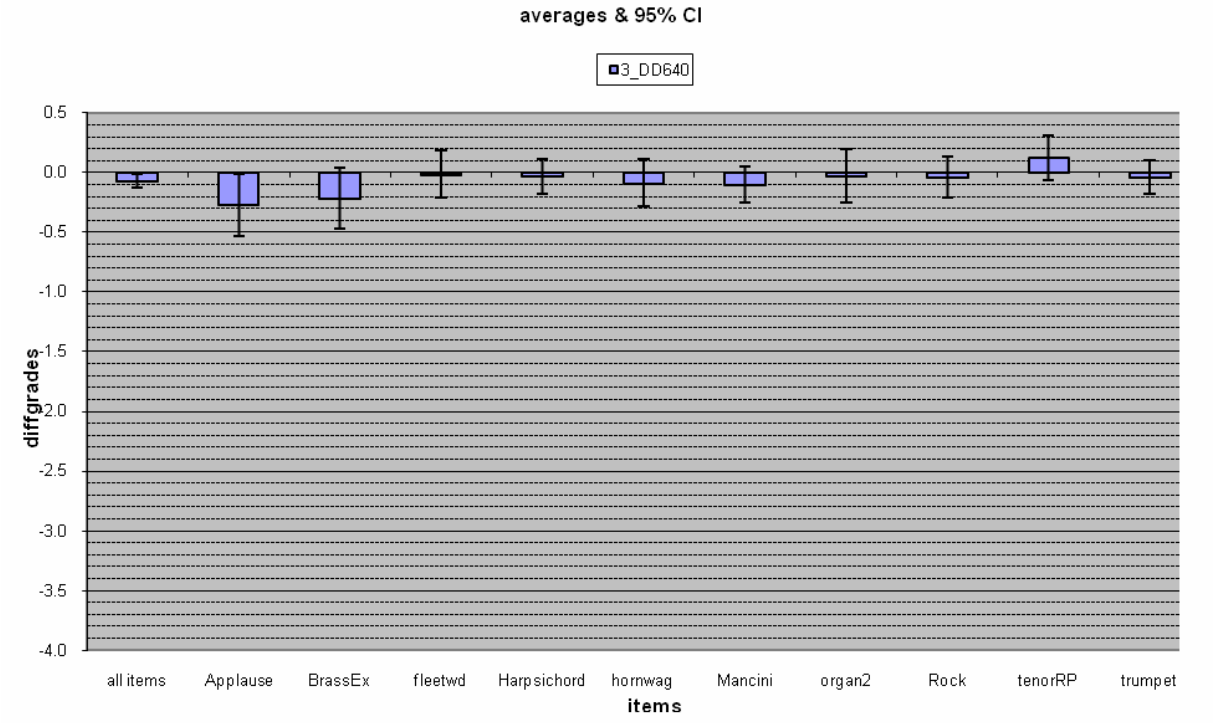
[5] ITU "Recommendation ITU-R BS.775-2: Multichannel stereophonic sound system with and without accompanying picture", July 2006

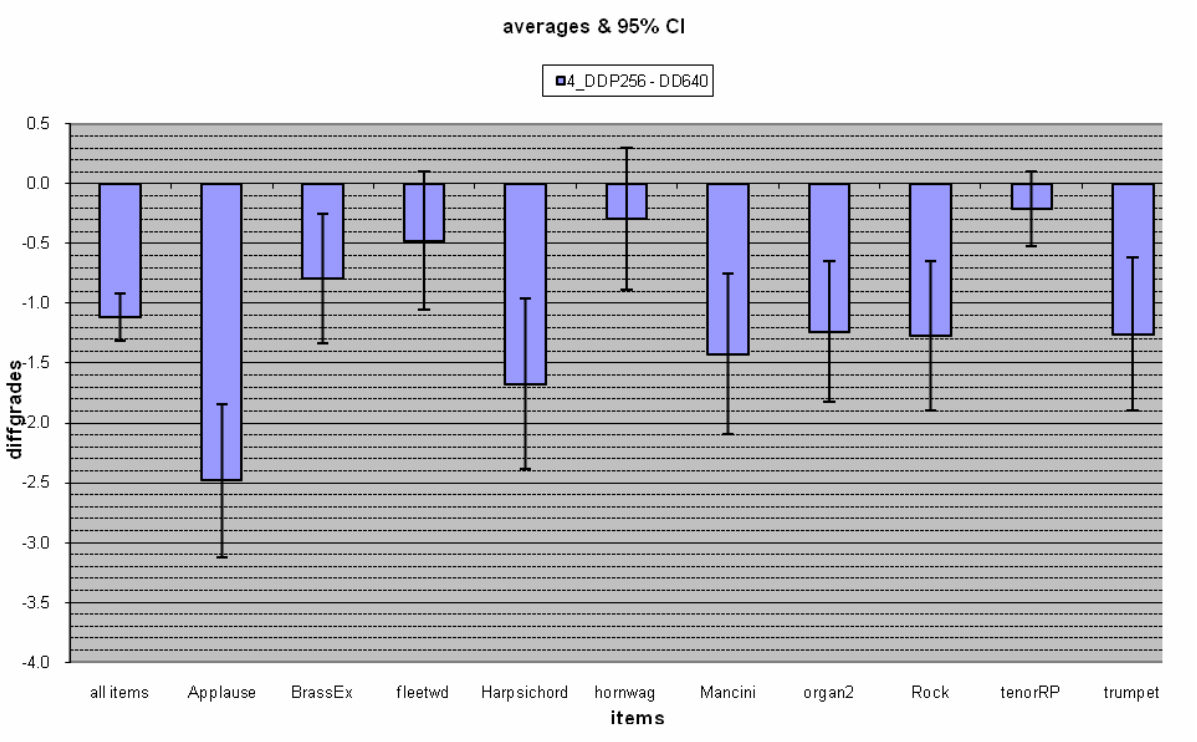
Appendix 1: Detailed test results of Phase 3

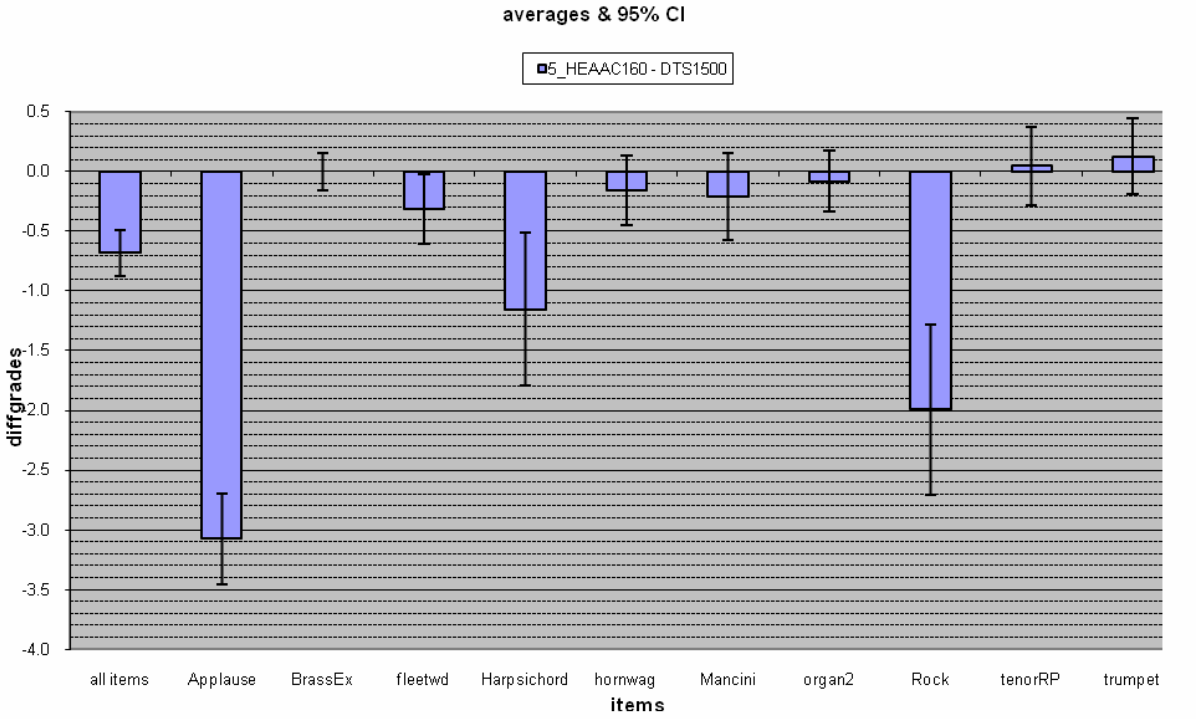
Averages for each codec over all test items

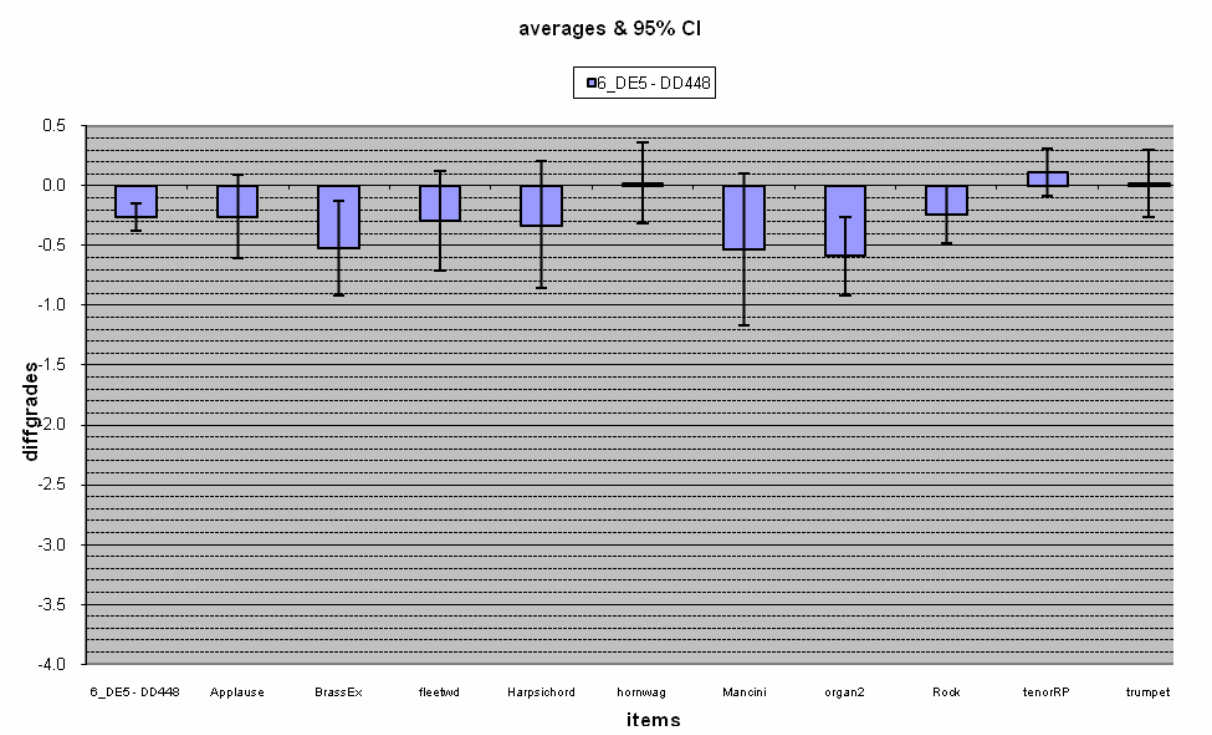
These graphs show how each codec scored over all items and for each of the test items. They highlight any particular strengths or weaknesses with particular types of material for each codec.

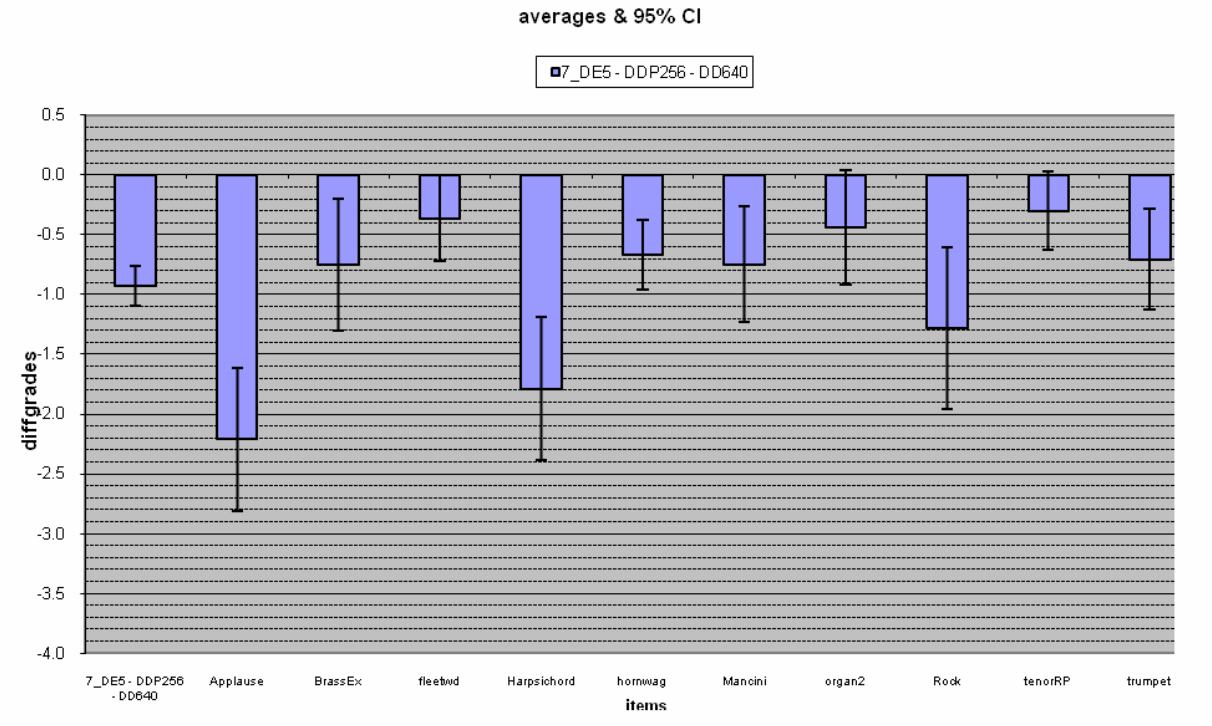


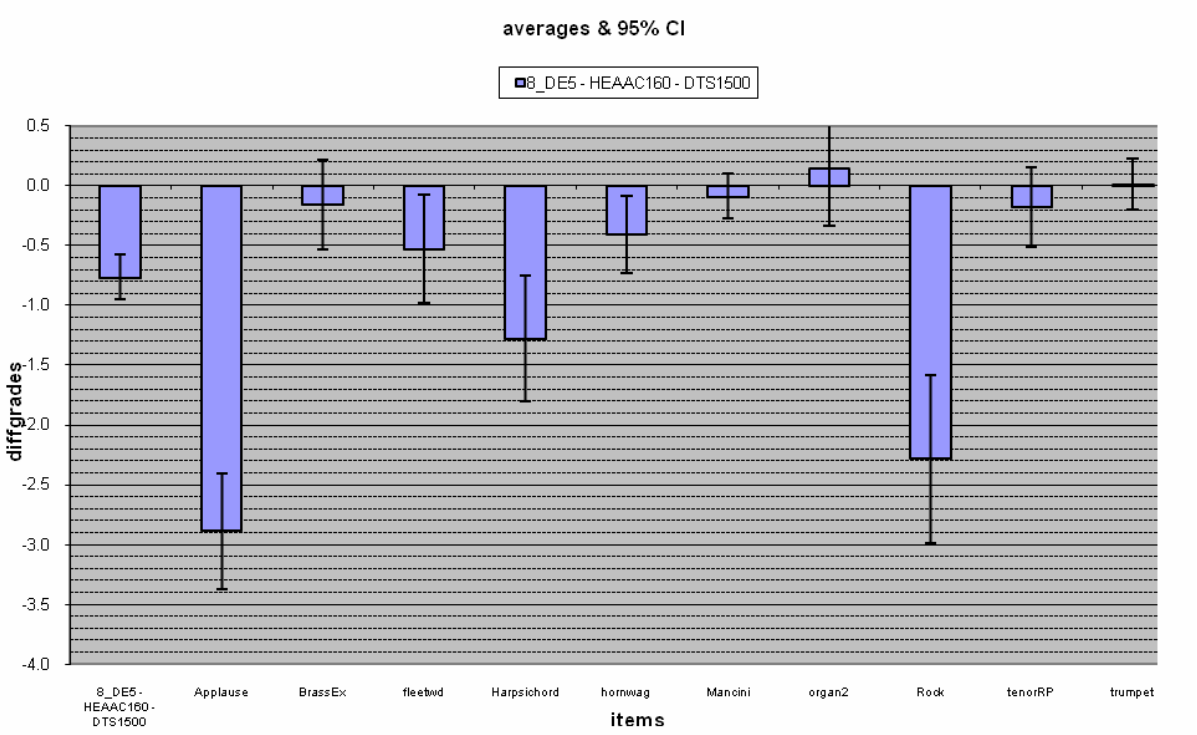


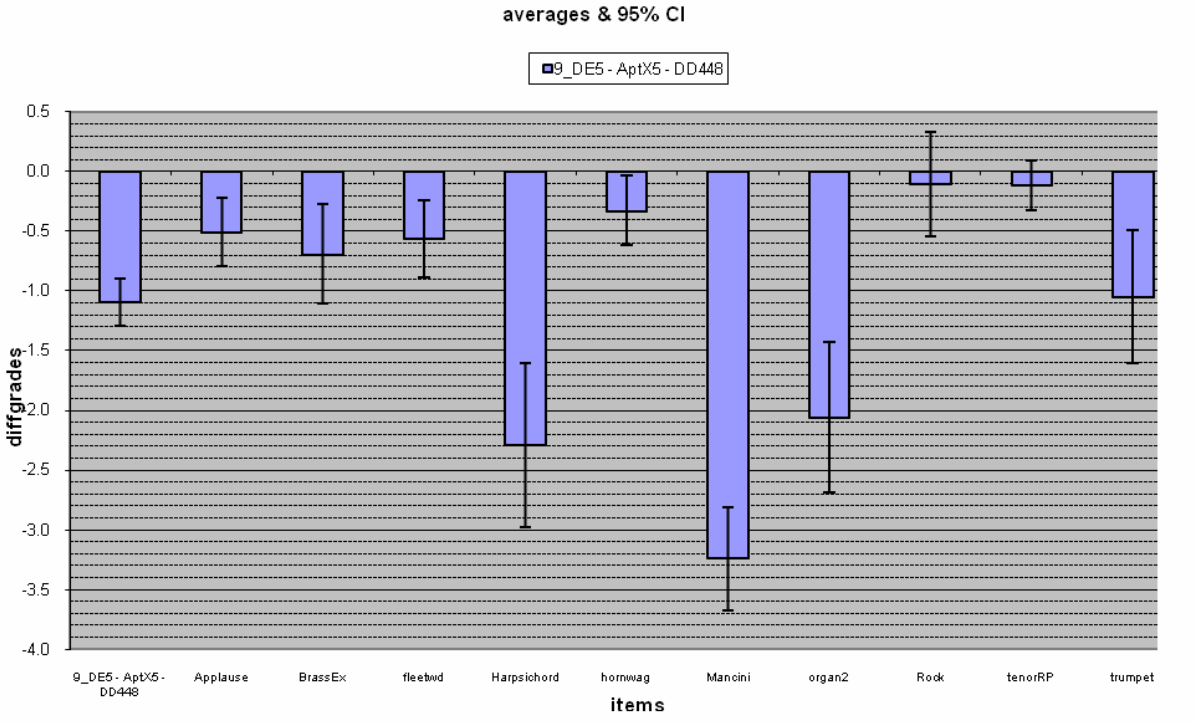

averages & 95% CI
4_DDP256 - DD640
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
diffgrades
all items
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items

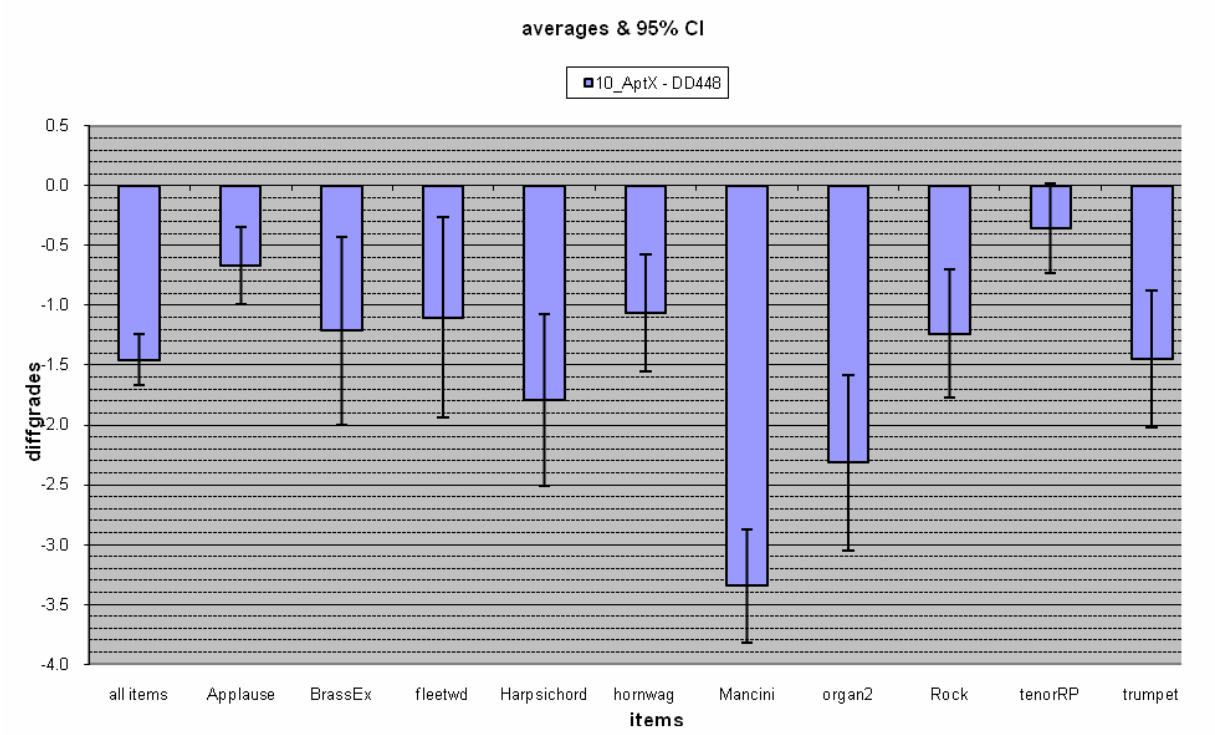

averages & 95% CI
5_HEAAC160 - DTS1500
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
diffgrades
all items
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items

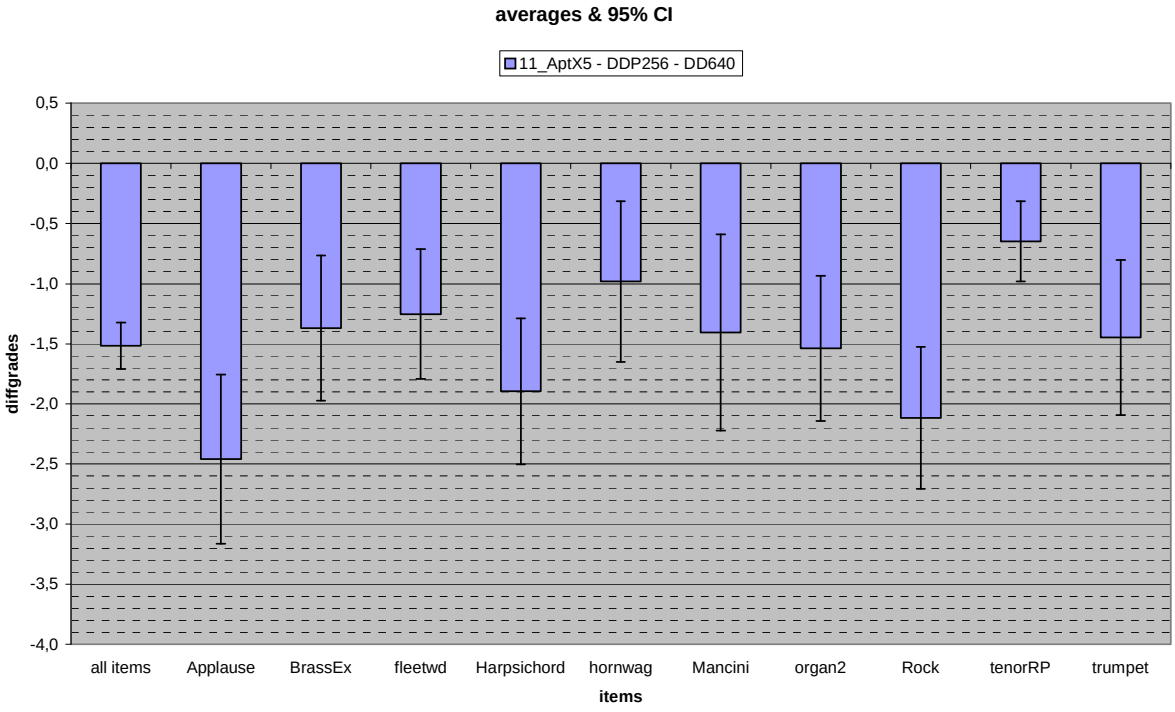

averages & 95% CI
6_DE5 - DD448
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
diffgrades
6_DE5 - DD448
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items

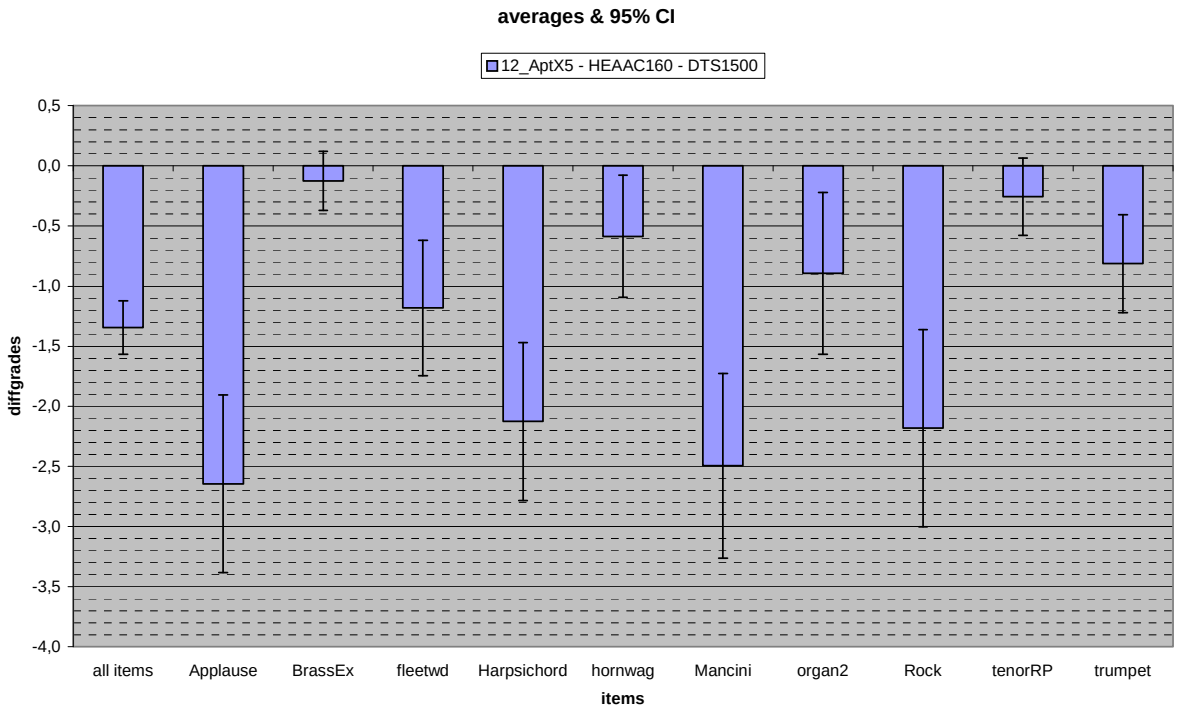

averages & 95% CI
7_DE5 - DDP256 - DD640
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
diffgrades
7_DE5 - DDP256 - DD640
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items

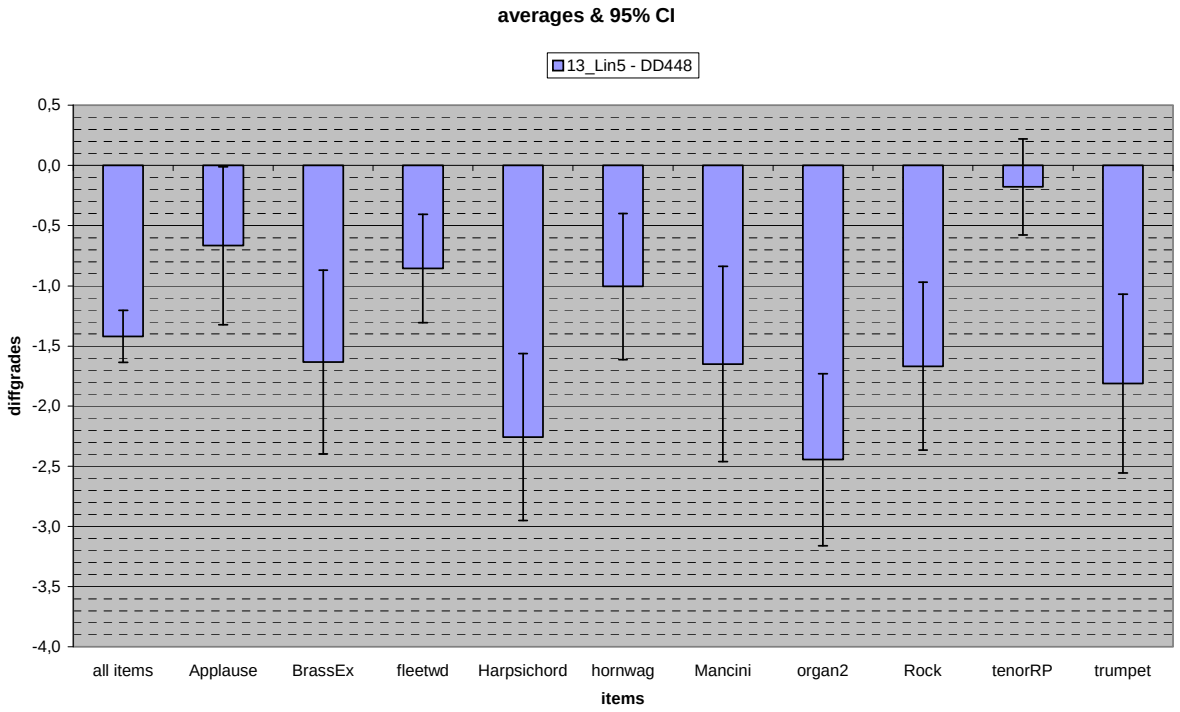

averages & 95% CI
8_DE5 - HEAAC160 - DTS1500
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
diffgrades
8_DE5 - HEAAC160 - DTS1500
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items

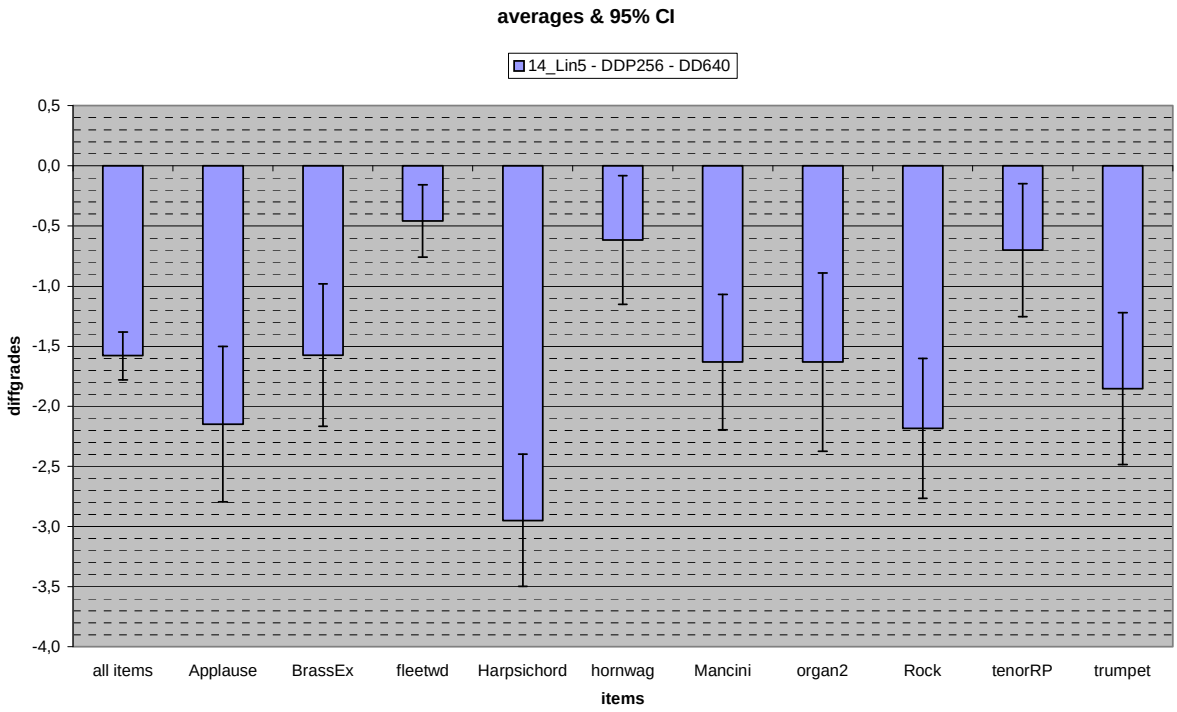

averages & 95% CI
9_DE5 - AptX5 - DD448
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
diffgrades
9_DE5 - AptX5 - DD448
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items

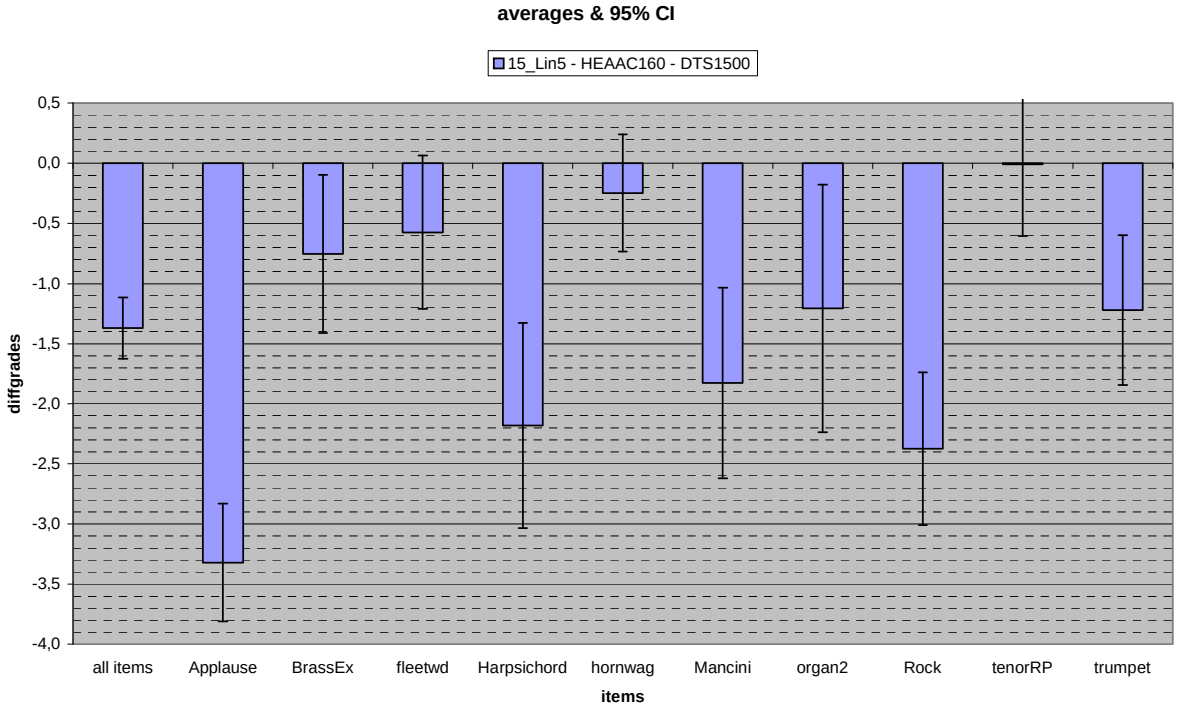

averages & 95% CI
10_AptX - DD448
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
diffgrades
all items
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items


averages & 95% CI
11_AptX5 - DDP256 - DD640
0,5
0,0
-0,5
-1,0
-1,5
-2,0
-2,5
-3,0
-3,5
-4,0
diffgrades
all items
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items


averages & 95% CI
12_AptX5 - HEAAC160 - DTS1500
0,5
0,0
-0,5
-1,0
-1,5
-2,0
-2,5
-3,0
-3,5
-4,0
diffgrades
all items
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items


averages & 95% CI
13_Lin5 - DD448
0,5
0,0
-0,5
-1,0
-1,5
-2,0
-2,5
-3,0
-3,5
-4,0
diffgrades
all items
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items


averages & 95% CI
14_Lin5 - DDP256 - DD640
0,5
0,0
-0,5
-1,0
-1,5
-2,0
-2,5
-3,0
-3,5
-4,0
diffgrades
all items
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items


averages & 95% CI
15_Lin5 - HEAAC160 - DTS1500
0,5
0,0
-0,5
-1,0
-1,5
-2,0
-2,5
-3,0
-3,5
-4,0
diffgrades
all items
Applause
BrassEx
fleetwd
Harpsichord
hornwag
Mancini
organ2
Rock
tenorRP
trumpet
items

[*] Page intentionally left blank. This document is paginated for two sided printing

Appendix 2: MCA Codec Descriptions

Technical details on the MCA codecs[1] used in the Phase 3 tests are as follows:

Enhanced apt-X

Standard apt-X, developed in 1990, compresses 16-bit PCM audio samples by 4:1. Enhanced apt-X was developed in 2000, for applications demanding higher quality audio.

Both algorithms are based on sub-band ADPCM. Standard apt-X takes a series of 16 bit PCM words and passes these through a 2-stage QMF tree, splitting the signal into 4 equally divided sub-bands.

In the sub-bands each 16 bit word is passed through a quantiser routine, this is then passed through an inverse quantiser and predictor circuit to predict the size of the next signal, which uses the history of a number of previous samples This prediction is compared with the actual signal and the "difference" is measured. Advanced mathematical principles are used to assign a value to this "difference" signal and this is what the encoder passes to the decoder. The decoding process is the inverse of the encoder.

Enhanced apt-X contains a number of modifications to the QMF, quantisers and predictors, to give a lower delay, higher audio resolution and a choice of bit depths (16, 20, 24 bit).

Latency through standard apt-X is due to 122 samples passing through the QMF stages and that equates to 2.5 ms at 48 kHz sampled audio. Within the Enhanced apt-X algorithm the QMF filters have been changed to more sophisticated coefficients which produce a delay of 90 samples which equates to 1.9 ms at 48 kHz sampled audio.

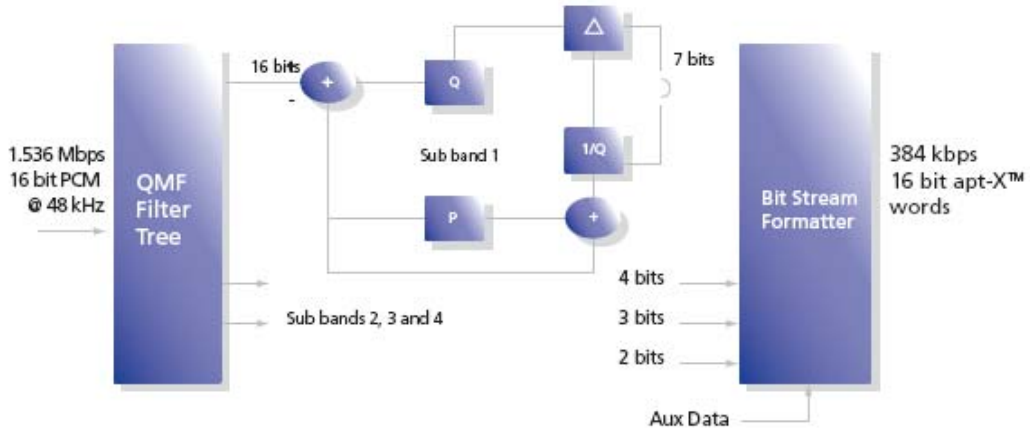


Figure A2-1: Enhanced apt-X encoder

[1] For the Linear Acoustic codec see the note below Table 3

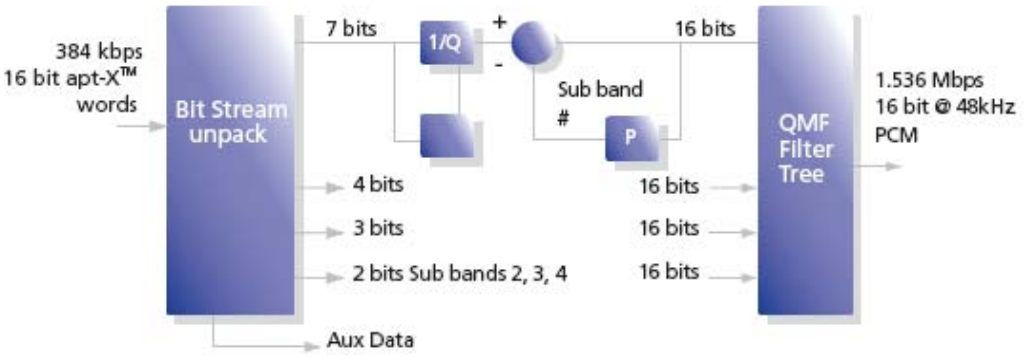


Figure A2-2: Enhanced apt-X decoder

Dolby E

Dolby E audio coding was developed specifically for the production, contribution, and distribution of multichannel audio. This is distinct from audio encoded with emission codecs such as Dolby Digital, which is used for final transmission of multichannel programs directly to the home viewer.

With Dolby E, coded audio frames match (that is, are aligned with) video frames, ensuring that audio-follow-video edits are free of mutes, glitches, or other aberrations. This makes it possible to switch, route, and perform edits directly on the coded bit-stream without decoding and re-encoding. The latency of the encode and decode process is two video frames (one frame for the encoder, one for the decoder). The Dolby E algorithm has been specifically designed to withstand the multiple encode/decode generations typically required during the production, contribution, and distribution phases of DTV. Dolby E audio also carries Dolby Digital metadata for final delivery to the home viewer's Dolby Digital decoder. A Dolby encoder will generate up to eight audio channels plus metadata and encode these into a single two-channel AES bit-stream (20-bit at 48 kHz). An alternate mode of operation allows for encoding of six audio channels plus metadata (16-bit at 48 kHz).

The audio channels within a Dolby E stream can be grouped together to carry separate audio programmes. For example, with multichannel programming, a "5.1 + 2" configuration is typically used, with six of the eight channels carrying a 5.1 channel programme and the other two a Lt/Rt (matrix surround-encoded) or stereo programme. Alternate configurations include a 5.1 programme plus two mono tracks (5.1 + 1 + 1), four stereo programmes (4 × 2), and eight mono channels (8 × 1).

Appendix 3: Members' Listening Rooms and Equipment Set-ups

The following is a description of the listening rooms and the equipment set-up of the EBU laboratories involved in the tests. The following laboratories contributed the relevant data[1]:

- Bayerischer Rundfunk (BR)
- Institut für Rundfunktechnik (IRT)
- Radiotelevisione Italiana (RAI CRIT)
- Sveriges Radio (SR)
- Westdeutscher Rundfunk (WDR)

Bayerischer Rundfunk (BR), Munich, Germany

Listening room:

Size:	Basic shape:	Mid-size room, acoustical prepared, nearly rectangular.
	Floor area:	50.4 m^2 (4.80 m x 10.50 m)
	Height:	2.55 m
	Volume:	~ 128 m^3
Acoustics:	Basic room treatment: panel absorbers, edge absorbers	
Loudspeakers:	5 x RL901K (Musikelectronic Geithain), 1 x BASIS 4 (Musikelectronic Geithain)	

[1] Data from BR and RAI could not be taken into account.

Technical equipment:

Audio workstation Hardware:	Fujitsu Siemens CELSIUS-Laptop, 2 GHz, 2 Gbyte RAM, Windows XP Professional
Audio interface:	RME Multiface II
Software:	STEP from Audio Research Labs
Additional Hardware/Software:	Stagetec I/O-Basedevice G&D-KVM-Extender for VGA and Mouse
Mixing-Console:	Stagetec AURUS-Console

Institut für Rundfunktechnik (IRT), Munich, Germany

Listening room:

Meets the following recommendations:

EBU Tech 3276 (including supplement 1 for multichannel monitoring)
ITU-R Recommendation BS.1116.

Loudspeakers:

5x MEG RL922 (Musikelectronic Geithain)
2x Genelec subwoofers (1092A + 1094A)

Technical equipment:

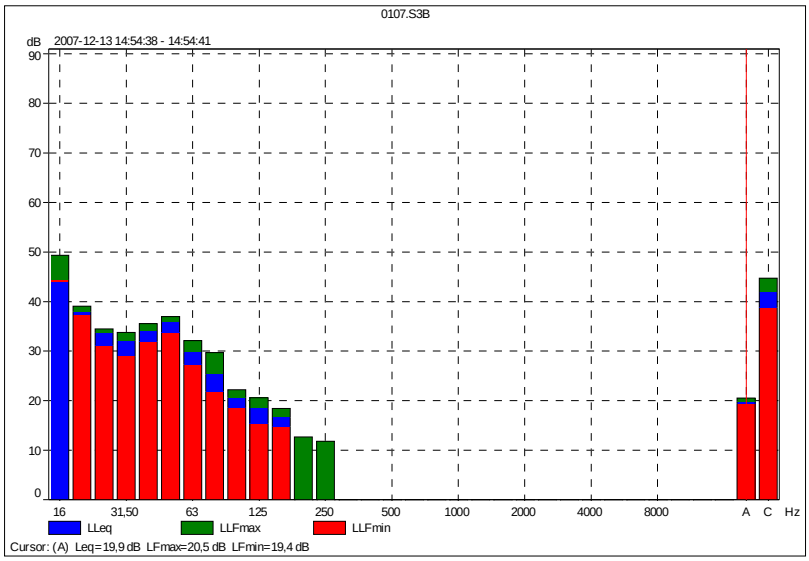
Mixing Console: Yamaha O2R
Additional Hardware/Software: Yamaha DAC

Sveriges Radio (SR), Stockholm, Sweden

Listening Room:

Floor area W & D:	6.0 x 6.50 m without absorbers/diffusers
Height:	3.06 m without absorbers/diffusers
Reference listening position:	Beam ring 2.2 m to the loudspeakers (as per ITU BS.775-2)
Acoustics:	Good isolation, both absorbers and diffusers, non-parallel walls and roof

N/C value (picture below):



Technical equipment:

Audio Workstation:	Windows XP Workstation with Lynx AES 16 card
Software:	ARL STEP, version 1.05
Loudspeakers:	Genelec 8020A with digitally controlled frequency response
Subwoofer:	Genelec 7270A (digital input)

Digital level was controlled by the DSP in the loudspeakers; the D/A converter is after the DSP in the Genelec loudspeakers

Centro Ricerche Innovazione Tecnologica (RAI CRIT), Turin, Italy

Listening room:

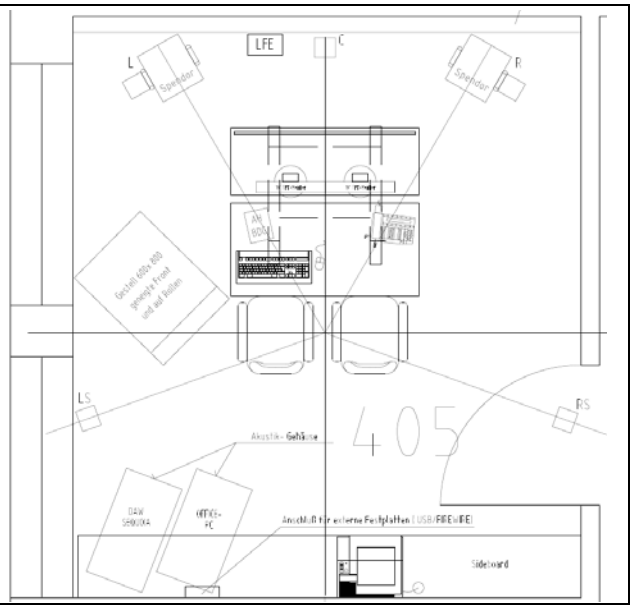
Size:	Basic shape rectangular
Acoustics:	Basic room treatment: drapery, diaphragmatic absorber
Floor area:	WxD = 500 x 800 cm
Height:	335 cm
Reference listening position:	centre of the ring beam 210 cm
Reverberation time/f	n/a
NC value	n/a
Loudspeakers:	Genelec 8050A monitors+ 7070A subwoofer
Frequency/phase response:	http://www.genelec.com/pdf/DS8050a.pdf http://www.genelec.com/pdf/DS7000_2.pdf

Technical equipment:

Audio workstation:	Windows XP Workstation with RME9652 digital interface
Software:	ARL STEP, version 1.05
Mixing Console:	Yamaha DM2000

Westdeutscher Rundfunk (WDR), Cologne, Germany

Listening room:



Basic shape:	Small room, acoustically treated, rectangular
Acoustics treatment:	panel absorbers, edge absorbers, NC value 15
Size:	12 m^2 (3.74 m wide, 3.26 m long)
Height:	2.66 m
Volume:	about 32 m^3
Reference listening position:	almost following ITU BS.775-2
Loudspeakers:	2 x Spendor 75/1A, 5 x Genelec 8030A, Subwoofer Genelec 7050B

Technical equipment:

Audio workstation:	Customized Windows XP PC, P4, 3.20 GHz with RME Hammerfall HDSP 9652, RME ADI 192 DD
Software:	Sequoia 9.1
Additional Hardware/Software:	RTW Surround Controller 30900