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

ISO/IEC JTC1/SC29/WG11
N2006
February 1998

Report on the MPEG-2 AAC Stereo Verification Tests

David Meares, BBC R&D, Kingswood Warren, UK
Kaoru Watanabe, NHK, Tokyo, Japan
Eric Scheirer, MIT Media Labs, USA

1.Table of Contents

2.Introduction

In November 1996, the details of the MPEG-2 Advanced Audio Coding (AAC) multichannel coding tests conducted at BBC and NHK were reported on in document WG11/N1419 [1]. Those tests showed a high quality performance for AAC at bit rates of approximately 320 kbps for 5-channel operation (64 kbps/ch). At that time, due to the differences in the use of a common bit reservoir and several joint processing strategies, the observation was made that although the 5-channel AAC had received a performance characterisation, it was still necessary to conduct separate tests of the stereo and mono performance of the AAC codecs.

As a consequence of that observation, the methodology and details of AAC formal stereo tests, both in the context of the established MPEG-2 (ISO/IEC 13818-7) standard and the forthcoming MPEG-4 (ISO/IEC 14496-3) standard were set forth in document WG11/N1845 [2].

Those tests have now been completed, and it is the purpose of this document to report the procedures, details and results of the tests.

3.Time schedule

The schedule of activities involved in the AAC stereo tests and the organisation conducting each phase of the work is listed below.

Activity	Deadline	Time	Responsible Company
Providing new excerpts	1 Aug. 97	1w	Teracom
Collecting and preparing new test material	8 Aug. 97	1w	Samsung
Delivering test material to proponents	15 Aug. 97	1w	Samsung
Coding excerpts, Delivering test material to verification site and pre-screening site	10 Oct. 97	8w	FhG, Sony, Philips. See Section 4
Verification of the encoded/decoded test materials	17 Oct. 97	1w	AT&T, Berkom, NSC. See Section 4
Selecting critical materials	24 Oct. 97	1w	
Site for selection			FhG
Critical listeners			BBC, Berkom Univ. of Hannover
Test Administration			AT&T
Delivering of critical material and bitstreams to the bitrate verification sites	7 Nov. 97	1w	FhG
Bitrate verification	14 Nov. 97	1w	Fivebats, BBC, AT&T
Strip sine burst	5 Nov. 97	10d	FhG
Deliver critical material to the test tape preparation site (AT&T)			

Activity	Deadline	Time	Responsible Company
Preparation of test and training tapes Deliver test and training tapes to the NHK test site	14 Nov. 97	10d	AT&T
Test set-up	14 Nov. 97	3w	NHK
Grading phase	5 Dec. 97	3w	NHK
Statistical analysis	19 Dec. 97	2w	MIT
Test report	18 Jan 98	2w	NHK, BBC, MIT

4.Codecs under test

Based on advice from the Selection Panel, the Audio Subgroup, at the October 1997 MPEG meeting in Fribourg, recommended the following codecs and bitrates be used in these tests.

Codec	Profile	Fixed Bitrate	Codec supplier	Independent check site and Comments
AAC	Main	96,128	FhG	AT&T
AAC	Low Complexity	96,128	FhG[1]	AT&T
AAC	SSR	128	Sony	NSC
Layer II		192	Philips	Berkom
Layer III		128	FhG	AT&T
codec_x		not to be identified		codec_x is used for some low anchor signals

During the material selection process, several stimuli from a codec referred to as codec_x were chosen in order to provide stimuli expected to give scores in the middle of the subjective range, i.e. approximately 2.5 on the impairment scale. These stimuli were to be included so that sufficient range of results could be ensured to facilitate checking the reliability of the listeners, see Section 10 below.

5.Test material

A call for new stereo test material was sent out after the MPEG meeting in Bristol, April 1997 [3]. This resulted in offers of 20 new test excerpts and these, together with the MPEG-1 original test excerpts, were used as the basic set for these tests. The full list of 42 items is given in Annex 1.

As with earlier subjective tests, the process of identifying and selecting the most critical programme items to be used in the formal tests was delegated to a selection panel. The selection panel was comprised of

- Andrew McParland, BBC R&D
- Thomas Buchholz, Deutsche Telekom, Berkom
- Lampos Ferekidis, University of Hannover

[1] In the original plan the low complexity profile was to be supplied by AT&T and the verification was to be done by FhG. This was changed to the conditions shown here during the execution of the test preparations.

and operated under the guidance of Jim Johnston as supervisor of the selection process. Their report, including the instructions given to them at the outset of their task, is presented in Annex 1.

The final selection of items used for the formal stereo tests is as follows.

Item No	File Name	Duration (sec)	Signal	Source
0	te3	16.6	Castanets	SQAM
1	te4	18.2	Harpsichord	SQAM
2	te5	28.4	Pitch Pipe	Dolby
3	te6	22.7	Glockenspiel	SQAM
4	te7	20.5	Male German Speech	SQAM
5	te8	21.0	Suzanne Vega, Tom's Diner	Solitude Standing
6	te9	29.5	Tracy Chapman	Elektra 960 774-2
7	te11	22.9	Ornette Coleman	Dreams 008
8	te16	19.9	Accordion/Triangle	Private (analogue) recording
9	te22	33.7	Dire Straits	Warner Bros. 7599-25264-2

It should be noted that, after the Selection Panel had reported its findings and after bitstreams had been supplied to the tape preparation site, it was decided that the recorded level for Glockenspiel was too high in relation to the general level of the other items. It was feared that this would cause discomfort or distraction to the listeners if the BS.1116 recommended line up procedure was adopted.

The MPEG Audio ad-hoc group for these tests, therefore decided to reduce the level for this test item by 0.5 (-6 dB) and to add 3/4 LSB triangular dither prior to re-quantisation. This was carried out for the source material and all the coded/decoded versions of this item. In other respects, the BS.1116 recommendation was to be adhered to as closely as possible.

6.Test methodology

The methodology used for these tests is based on the ITU-R Recommendation BS.1116 [4], the triple stimulus/ hidden reference/ double blind methodology. Most of the details of that methodology were adopted, as were the constraints on room acoustics etc., except as mentioned here.

BS. 1116 specifies that for the greatest listener sensitivity to artefacts each listener should be tested on his/her own and should be free to switch at any time between the stimuli under assessment. The current tests, however, had to be conducted under severe time constraints, and it was necessary to ask up to three listeners at a time to participate simultaneously. This meant that it was not possible to allow listener controlled switching. As a result, a pre-recorded sequence of stimuli Ref/A/B/Ref/A/B were recorded on tape as is described in Section 7.1.

Additionally, as these were stereo performance comparisons, the seating arrangements of BS. 1116 were modified, as also detailed below, to allow for the multiple listeners per test session.

In summary, the conditions applying to these tests were as follows:

- triple stimulus/hidden reference/double blind method
- pre-recorded test material on DAT tape in Ref/A/B/Ref/A/B arrangement
- BS.1116 attribute "Basic Audio Quality" for grading

- five grade impairment scale, see Section 6.4
- one of the stimuli ‘A’ or ‘B’ must be graded with 5.0 (hidden reference)
- loudspeaker arrangement (as shown in Annex 2)
- listeners to be beyond the critical distance (if possible, see below)
- a maximum of 3 subjects at a time to participate
- a minimum of 20 expert listeners
- fixed presentation of SPL (no adjustment by the subjects)

6.1. Listening conditions

The listening room at NHK fulfils most of the requirements of BS.1116 and has been successfully used in similar tests. The geometric details of the listening room and the relevant features of the acoustics of the room and the loudspeakers are given in Annex 2.

One of the requirements of the test specification was that the listeners should make their evaluation ‘beyond the critical distance’. The justification for this was the experience that some of the more important stereo imaging artefacts, witnessed by the Selection Panel, are only audible under such conditions. The critical distance from a sound source in a room is defined as the distance from the source at which the direct sound level from the source is equal to the reverberant sound level due to that source. Mathematically this is given by the equation

$$\frac{Q}{4\pi r^2} = \frac{4(1-\alpha)}{S\alpha}$$

where
r = critical distance
Q = directivity factor of the loudspeaker
α = room absorption coefficient and
S = room surface area.

Additionally α may be computed as follows

$$\mathrm{T} = -0.162 V / S \lg_e (1-\alpha)$$

where
T = room reverberation time and
V = room volume

For the conditions prevailing in the NHK listening room, and assuming, as an approximation, that the directivity factor of the loudspeakers varies linearly from 1 at 50 Hz (i.e. omni-directional) to 4 at 10 kHz, the critical distance for the NHK room is given in Figure 1.

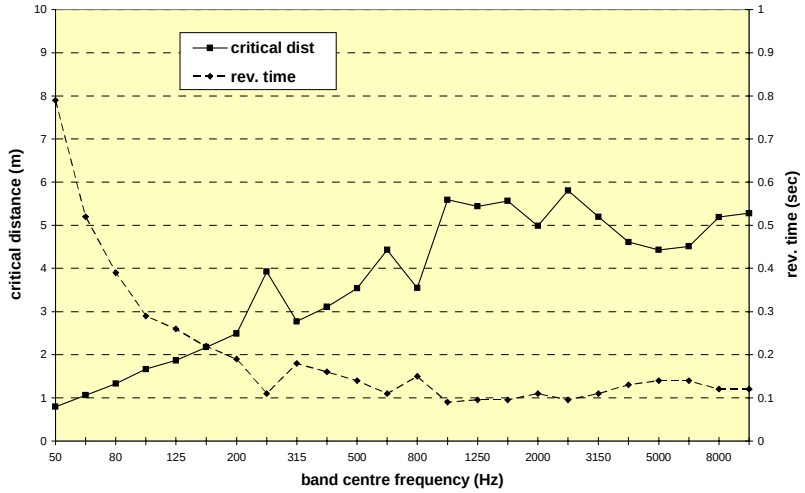


Figure 1 Critical distance in NHK listening room

As can be seen from this estimation of the critical distance, the listening positions used in these tests are in the transition area, being beyond the critical distance at low frequencies and within the critical distance at high frequencies. This is one of the unavoidable consequences of such a short reverberation time.

6.2. Listening position

Annex 2 shows the three listening positions at distances of 2.3m, 3.2 m and 4.15 m from the circular arc between the stereo loudspeakers. Thus the angle subtended by the loudspeakers at each of the three listening positions was 60^{o}, 42^{o} and 32^{o} .

6.3. Test equipment

The loudspeakers used for testing were high quality studio monitors, Mitsubishi type 2S-3003. The listening level at the Reference listening position was adjusted to give an SPL of 82 dB(A) for each loudspeaker, by means of a pink noise signal with the same RMS value as a 1 kHz tone at -18 dBFS (in accordance with BS.1116).

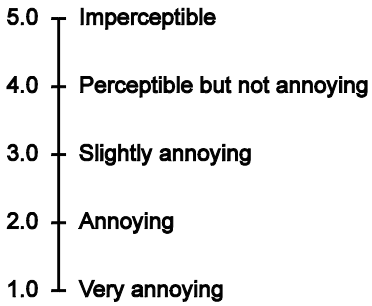
Once aligned, the two loudspeakers matched to 1 dB or better, both wideband and narrowband.

The digital devices used in these tests all had a resolution of 16 bits or more. All decoded audio passages were rounded as appropriate to 16 bit accuracy.

6.4. Grading

The listeners were asked to judge the single attribute ”Basic Audio Quality“ (BAQ), as proposed in the ITU-R Recommendation BS.1116. BAQ includes all audible differences between the reference and the coded version.

The listeners were instructed that the grades for the tests were to be given according to the ITU-R 5-point impairment scale as shown alongside. This was described to them as a continuous scale with anchor points. In awarding their grades to stimuli A and B, the subjects were required to grade at least one of A or B as 5.0 (the one they judged to be the hidden reference) and to give their results to one decimal place.



7.Preparation of the test material

7.1. Test stimuli

In triple stimulus/hidden reference/double blind method, three audio stimuli, reference "Ref", signal "A", and signal "B", are assessed by the listeners. "Ref" and one of the audio signals "A" or "B" are the reference or uncoded source material, whilst the remaining stimulus "B" or "A" is the coded material. The allocation of "A" and "B" to the hidden reference or the coded version is decided at random and the identity is known by neither the listener nor the person running the test.

The stimuli were, therefore, grouped into a number of test sequences or trials pre-recorded onto tape in the sequence shown in Figure 2. The progress of the tests was conveyed to the listener by means of announcements recorded in a sequence; i.e. "Item N", "R", "A" and "B.

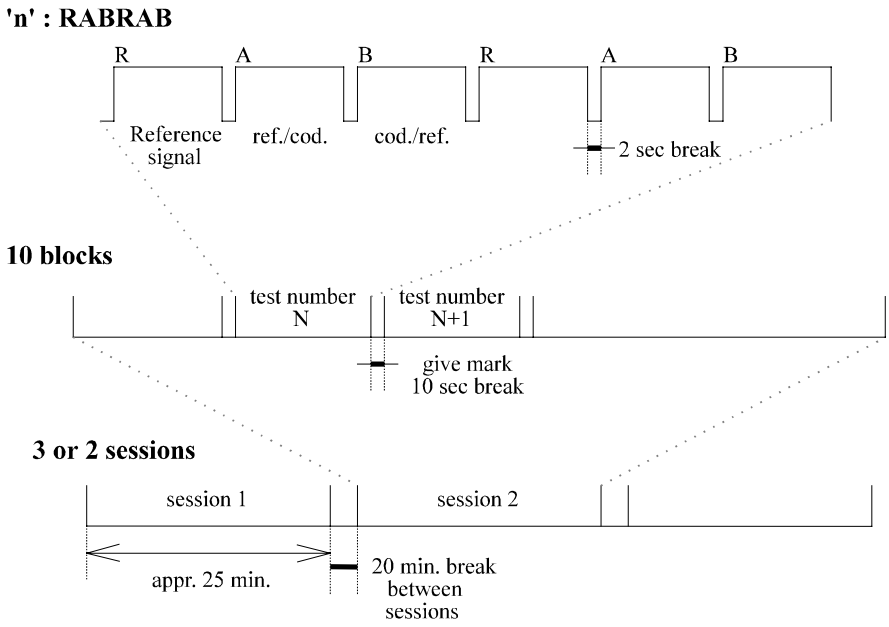


Figure 2. Protocol of triple-stimulus, hidden-reference, double blind test method

7.2. Test session duration

To ensure fatigue did not affect the results, each test session was restricted to a maximum of 25 to 30 minutes in duration. Also, between consecutive sessions listeners took a break of at least 20 minutes duration. The tests required a total of eight sessions per listener which they completed normally over a two day period.

7.3. Preparation of test blocks (randomisation)

Randomisation of the test stimuli was applied to minimise the number of times each codec configuration and each test item occurred in a test session and also to take into account the quality of the coded versions, i.e. there was a mix of audio quality throughout the test.

7.4. Test tape preparation

The stimuli used for the training session and the grading sessions were recorded in the described manner onto DAT tapes by AT&T and were sent to the NHK test site.

8.Listening panel

8.1. Subjects

To achieve statistically reliable results, 31 listeners participated, see Annex 3. They all had a background in professional audio work. Codec developers involved in the AAC tests were excluded as listeners to avoid any chance of bias, real or imagined.

There were no audiometric tests. Subject reliability was to be verified by post-processing of the results.

8.2. Training of the subjects

In order to train the subjects prior to the tests, a training session was used. During that time, some coded excerpts were replayed at various bitrates to show the range of artefacts which were present. There was guidance and support during training and testing from the test site personnel. The subject training followed the outline given below:

a) General introduction to the tests.
b) All reference versions of the test items were listened to, in order to get accustomed to stereo sound and to the test items themselves.
c) The 4 training excerpts (a subset of the ten test excerpts), coded with one of the codecs under test, were listened to and the perceived artefacts were discussed between subjects, under guidance but without talking about grades.
d) Where appropriate, difference signals, i.e. reference minus coded, were replayed to guide the listeners to the points where artefacts may occur.
e) This step repeated the c) and d) steps. The same training excerpts, coded with the other codecs under test, were listened to and the perceived artefacts were discussed.
f) The training was finished with a dummy test, using a few items. As with the main tests, each listener scored their test individually although the results were discarded. It was just for listener familiarisation.

9.Independent coder checks[2]

9.1. Bit rate verification

The task of bit rate verification was undertaken as follows:-

---

2 Reported below are the results of verifications on only those coders and bitrates included in the final tests. Additional coder/bitrate combinations that were submitted to the Selection Panel were also checked and confirmed.

Codec	Independent check site
AAC main	FiveBats
AAC lc	FiveBats
AAC SSR	FiveBats
Layer II	BBC
Layer III	AT&T

On behalf of FiveBats, Mr. Coleman reported that he had tested the AAC bitstreams for bitrate verification and found:

- All SSR bitstreams have a bitrate within 1% of nominal.
- Main and LC bitstreams at 128 kbps have bitrates within 1% of nominal.
- Main and LC bitstreams at 96 kbps have bitrates within 1% of nominal.

Mr. Quackenbush, for AT&T, reported that he checked the bitrate for all coders to be tested in the AAC stereo test and found that the rates were no more than 0.5% from nominal for all coders.

- aac_main/96 average rate is 96.462 kbps or 0.5 % from nominal
- aac_main/128 average rate is 128.635 kbps or 0.5 % from nominal
- aac_lc/96 average rate is 96.434 kbps or 0.5 % from nominal
- aac_lc/128 average rate is 128.636 kbps or 0.5 % from nominal
- aac_ssr/128 average rate is 128.720 kbps or 0.5 % from nominal
- layer2/192 average rate is 191.884 kbps or 0.0 % from nominal
- layer3/128 average rate is 128.243 kbps or 0.2 % from nominal

For the BBC, Mr. McParland reported that he had looked at the 10 MPEG-1 Layer II bitstreams. The bitrate indicated in the bitstreams was correct. The number of bytes per frame was correct for the indicated bitrate and did not vary through the files. The files were additionally decoded by a public domain MPEG decoder and sounded OK.

Thus, it can be concluded that the checks of coded bitrate all confirmed it to better than 1% of nominal.

9.2. Encoder/Decoder check

The task of encoder/decoder verification was undertaken as follows:-

Codec	Independent check site
AAC main	AT&T
AAC lc	AT&T
AAC SSR	NSC
Layer II	Berkom
Layer III	AT&T

For AT&T, Mr. Quackenbush reported his check of the encoders and decoders as follows.

a) AAC Main Profile Encoder Verification

I have encoded the following signals at the listed bitrates

signal	bitrates	
te5	128	96
te10	128	96
te15	128	96

and find that the AAC Main Profile encoder supplied by FhG produces bitstreams that exactly match the bitstream files supplied by FhG.

b) AAC Main Profile Decoder Verification

I have decoded the following bitstream files at the listed bitrates

signal	bitrates	
te5	128	96
te10	128	96
te15	128	96

and find that the AAC reference decoder produces PCM output files that exactly match the PCM output files that were supplied by FhG.

c) MPEG-1 Layer III Encoder Verification

I have encoded the following signals at 128 kbps

te5
te10
te15

and find that the Layer III encoder supplied by FhG produces bitstreams that exactly match the bitstream files supplied by FhG.

d) MPEG-1 Layer III Decoder Verification

I have decoded the following bitstream files

te5
te10
te15

and find that the Layer III decoder produces PCM output files that exactly match the PCM output files that were supplied by FhG after stripping the sine burst.

Mr Fukuchi, Nippon Steel, carried out the verification of encoded/decoded test material for the SSR profile. NSC received the original PCM files, Sony encoded SSR profile bitstreams, Sony decoded PCM files, and SSR profile encoder/decoder software for a SUN workstation from Sony. They encoded all the material, 42 items, at 128 kbps and decoded all the materials at same bitrate. They confirmed that all the test materials from Sony were identical to those which they produced independently by using Sony provided software.

Mr. Feige of Deutsche Telekom, Berkom, reported that the bitstream verification for MPEG-1 Layer II was successful. The bitstreams and decoded audio files obtained with the Philips coder were identical with the files provided on the CD-ROM. Additionally he decoded the bitstreams with a decoder from the CCETT and obtained files with sample differences of one LSB, maximum.

Thus, it can be concluded that the independent checks on coder/decoder validity were all positive.

9.3. Bitstream exchange

In order to conduct and verify these tests, a large number of audio bitstreams had to be created and exchanged between development and test sites. In total, 378 bitstreams were made available to and were decoded by FhG, the site conducting the selection panel work. In additional, over 240 bitstreams were exchanged with the various sites conducting the bitrate checks and the encoder/decoder verifications.

These confirm the interchangeability of AAC stereo bitstreams produced according to the standard.

10.Statistical analysis

The aim of the analysis was to answer the following questions, with supporting graphical presentations.

Based on these test results,

- Are the listeners' results reliable, i.e. distinguishable from random votes?
- Does the test methodology allow meaningful conclusions to be drawn from these results?
- Is the performance of AAC codecs at the tested bitrate equal to or better than the performance of MPEG-1 Layer II and Layer III?
- How does the performance of the codecs vary with programme items?
- Is the performance of the coding of AAC codecs at the tested bitrate distinguishable from the original signal?
- Is the performance of AAC codecs at the tested bitrate achieving 'indistinguishable quality' in the EBU definition [5] of that phrase?
- Is the following requirement of ITU-R Recommendation BS.1115 [6] fulfilled? "For emission, the most critical material for the codecs must be such that the degradation may be 'perceptible but not annoying' (grade 4)"
- What is the relative ranking of the codecs tested?
- Are there any other features from the data that should be reported?

10.1. Data receipt and organisation

The statistical analysis was carried out by Mr. Eric D. Scheirer, MIT Media Laboratory over the period 5 Dec. – 19 Dec. 1997. Data were received by MIT from NHK on 12 Dec. 1997 and randomised tape index information had been received from AT&T on 6 Dec. 1997.

The data were provided as an Excel spreadsheet. These data were rewritten in ASCII form, tab-delimited, to a temporary file. A PERL script (see Annex 4) was used to unroll the data into item-by-item lines, and to unblind or de-randomise and restore coder identities in the data with the use of the index information.

The resulting data file has 2480 lines (31 subjects x 80 stimuli/subject). These data were imported into SPSS V7.5S for Windows 95 for analysis; except as noted, all analysis was conducted with this tool. The data columns from this file correspond to the following SPSS variables:

SUBJ :	the subject number
SEAT:	the listening position of the subject
ITEM :	the critical sound example for the test case
CODER:	the coder used for the test case
REF:	the subject's rating of the reference excerpt
TEST:	the subject's rating of the test excerpt

Each row (or "case") in this file corresponds to one instance of one listener hearing and rating one critical excerpt as altered by one coder.

The correctness of the data unrolling was assessed in several ways. First, the independent variables were tabulated. There were 8 values for CODER[3], with 310 cases for each. There were ten values for ITEM, with 248 cases (31 listeners x 8 coders) for each. There were 31 values for SUBJ, with 80 cases for each. There were 3 values for SEAT, with 880 cases (11 listeners x 80 trials) for positions 1 and 2, and 720 cases (9 listeners x 80 trials) for position 3. Each of these is correct.

Cross-tabulations were computed. The ITEM x CODER matrix was size 10 x 8, with 31 cases in each cell. The CODER x SUBJ matrix was size 8 x 30, with 10 cases in each cell. The ITEM x SUBJ matrix was size 10 x 30, with 8 cases in each cell. Each of these is correct.

Diffscores (see [1]) were computed for each case and recorded as variable DIFF. For each diffscore, a negative value indicates the amount of impairment judged by the listener for that stimulus. Larger negative values indicate more impairment. A positive diffscore value indicates that the listener misperceived which was the test, and which the reference, signal for that case.

The mean diffscore for the data was -.5387, which is consistent with the data being unblinded properly. If the test and reference data values were randomised, the mean diffscore would be nearly 0; if they were exchanged, the mean diffscore would be positive. A histogram plot (see Figure 3) gives the expected shape for the diffscore distribution. At this point, the unrolling and unblinding were assumed correct and formal analysis was begun.

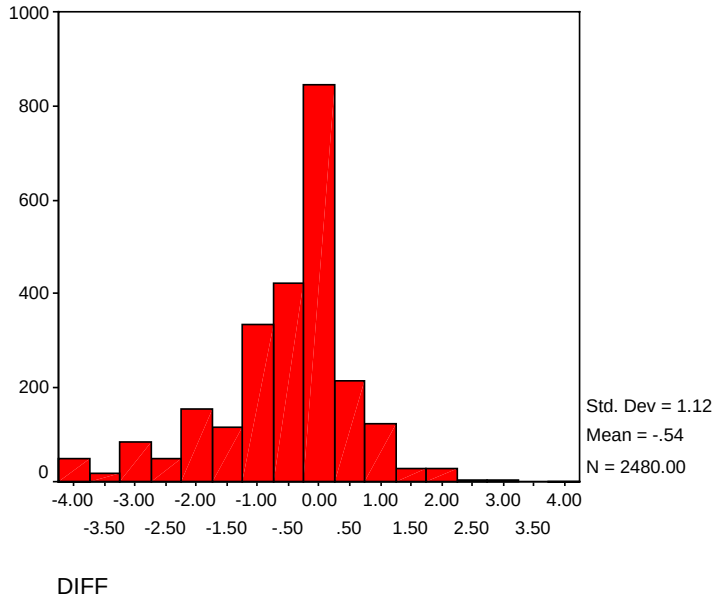


Figure 3. Diffscore histogram. The distribution of diffscores (left-skewed, centrally-clustered, mean slightly less than 0) is evidence that the randomised data were unscrambled properly.

10.2. Subject reliability

Subject reliability was assessed by ensuring that each subject gave mean diffscores which differed significantly from 0 in the negative direction. If the mean response was 0 or positive

[3] 7 coders for these assessments and a further set of stimuli, labeled codec_x, to ensure an adequate range of qualities.

for a particular subject, it would indicate that that subject was unable consistently to distinguish reference from test signals. According to the agreed test-protocol, subjects who are not reliable in this sense are to be removed from the data set.

The following table shows the results of a t-test applied to each subject's data. As can be seen, each subject has negative mean DIFF, and this value is significantly different from 0 at the $p < 0.05$ level. (The significance scores shown are two-tailed values, and so only need be less than 0.1 in order to reject the one-tailed null hypothesis that the listener is unreliable for this task). No listeners are rejected on this basis for this data set.

SUBJECT	t	df	Sig	Mean DIFF	Lower	Upper
0	-4.740	79	.000	-.4988	-.7082	-.2893
1	-4.591	79	.000	-.4875	-.6989	-.2761
2	-3.924	79	.000	-.5213	-.7856	-.2569
3	-3.522	79	.001	-.4275	-.6691	-.1859
4	-5.733	79	.000	-.6250	-.8420	-.4080
5	-4.725	79	.000	-.7687	-1.0926	-.4449
6	-5.519	79	.000	-.9463	-1.2875	-.6050
7	-5.337	79	.000	-.5600	-.7689	-.3511
8	-4.191	79	.000	-.5688	-.8388	-.2987
9	-7.145	79	.000	-.8587	-1.0980	-.6195
10	-3.000	79	.004	-.5475	-.9107	-.1843
11	-3.515	79	.001	-.2863	-.4483	-.1242
12	-4.652	79	.000	-.5000	-.7139	-.2861
13	-3.995	79	.000	-.3375	-.5057	-.1693
14	-4.762	79	.000	-.1988	-.2818	-.1157
15	-7.835	79	.000	-1.2188	-1.5284	-.9091
16	-4.421	79	.000	-.6650	-.9644	-.3656
17	-3.786	79	.000	-.5525	-.8430	-.2620
18	-2.456	79	.016	-.2687	-.4866	-.0509
19	-4.228	79	.000	-.6313	-.9284	-.3341
20	-2.014	79	.047	-.2013	-.4002	-.0023
21	-3.472	79	.001	-.2088	-.3284	-.0891
22	-5.198	79	.000	-.4675	-.6465	-.2885
23	-3.469	79	.001	-.2575	-.4052	-.1098
24	-4.720	79	.000	-.5425	-.7713	-.3137
25	-3.984	79	.000	-.5550	-.8323	-.2777
26	-4.804	79	.000	-.6713	-.9494	-.3931
27	-5.753	79	.000	-.8763	-1.1794	-.5731
28	-5.195	79	.000	-.8063	-1.1152	-.4973
29	-3.480	79	.001	-.3013	-.4736	-.1289
30	-4.214	79	.000	-.3425	-.5043	-.1807

10.3. Effect of listener position

For this test, two or three listeners were simultaneously presented with the stimuli, seated in a room as described in the test protocol. It is necessary to assess the effect of the different listener positions on the result, in order to test whether it is appropriate to pool the results for all listeners in subsequent analyses.

A one-way ANOVA was calculated to examine the effect of listener position on DIFF. The result shows a significant ($p=0.011$) influence of listener position on diffscore.

		Sum of Squares	df	MS	F	Sig.
DIFF	Between Groups	11.269	2	5.635	4.507	.011
	Within Groups	3096.632	2477	1.250		
	Total	3107.902	2479			

Post-hoc tests (the Tukey matrix) were employed to examine the nature of this influence.

(I) SEAT	(J) SEAT	Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
1	2	5.091E-02	.053	.605	-.0740	.1758
	3	-.1153	.056	.100	-.2470	.0163
2	1	-5.0909E-02	.053	.605	-.1758	.0740
	3	-.1662*	.056	.009	-.2979	-.0346

* The mean difference is significant at the .05 level.

As can be seen from this table, listeners in the third position had significantly (p = 0.009) less negative DIFF than listeners in the second position, indicating that they were significantly less able to identify artefacts in the test signals. Additionally, there was a trend which was not significant (p=0.1), but in the same direction, between the listeners in the first position and listeners in the third position. There were no statistical differences between listeners in the first and second positions.

These differences are shown graphically in Figure 4. As can be seen, the 95% confidence intervals do not overlap for position 2 and position 3, and barely overlap for position 1 and position 3.

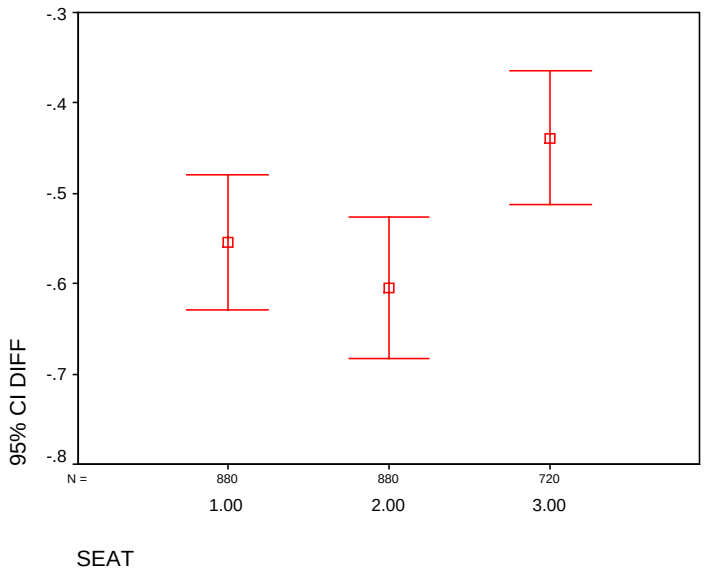


Figure 4: One-way analysis of variance, showing effect of seat position on mean diffscore, pooled for all items and codecs. The mean diffscore for position 3 is significantly less negative than for position 2, indicating that the differences between codecs were less perceptible from this listening position.

This result means that it is improper to pool results from listeners in positions 2 and 3, and questionable for listeners in groups 1 and 3. Thus, the main part of the further analysis was conducted on the 22 listeners in positions 1 and 2, pooled for analysis. At a later date, the data from listeners in position 3 may be analysed and contrasted with data from the main group, but with the small number of listeners in this group, there is little statistical power available.

Further analysis of the main group (excluding listeners in position 3) showed continuing weak influence of listening position on results. In particular, although there was no main effect of

position, there was a weak interaction effect ($p = 0.05$) between ITEM and SEAT, suggesting that the effect of SEAT differs depending on which ITEM is presented. There was no interaction between SEAT and CODER.

It is a difficult question whether it is still proper to pool results from position 1 and 2 given this weak interaction. To separate the data would mean that the largest single pool of subjects would have only 11 listeners; a group this small has limited statistical power. Further, since the primary variable of interest is the differences among the coders, and the position does not interact with this variable, positions 1 and 2 were kept pooled together.

10.4. Evaluation of coders

A two-way ANOVA was conducted to examine the effects of CODER and ITEM on the main listener group's results.

		Sum of Squares	df	MS	F	Sig.
Main Effects	CODER	158.540	7	22.649	26.447	.000
	ITEM	458.835	9	50.982	59.532	.000
2-Way	CODER * ITEM	309.771	63	4.917	5.742	.000
	Model	927.145	79	11.736	13.704	.000
	Residual	1438.718	1680	.856		
	Total	2365.864	1759	1.345		

This result shows that there is a main effect of the codec used (i.e., some codecs sound better than others), of the stimulus item (i.e., some items mask coding artefacts better than others), and an interaction between the codec and item (i.e., some items reveal particular artefacts in different coders). All of these effects are highly significant ($p < 0.001$).

The table in Annex 5 shows item-by-item and coder-by-coder breakdown of the means and confidence intervals of the diffscores for the various coders. Graphs of each of these results is shown in Annex 6.

There are several questions posed by the test protocol which can be answered using these results.

10.5. Differences between programme items

First, "how does the performance of codecs differ by programme item?" We will consider each of the AAC codecs in turn, comparing the confidence interval of the diffscores for that codec for each item to the MP2 and MP3 results. If the confidence intervals do not overlap, we judge one coder to be better for that item.

AAC Main 128:

Better than MP2 for 3 items, worse for no items, equivalent for 7 items.
Better than MP3 for 3 items, worse for no items, equivalent for 7 items.

AAC Main 96:

Better than MP2 for 1 item, worse for 1 item, equivalent for 8 items.
Better than MP3 for 1 item, worse for no items, equivalent for 9 items.

AAC LC 128:

Better than MP2 for 3 items, worse for no items, equivalent for 7 items.
Better than MP3 for 3 items, worse for no items, equivalent for 7 items.

AAC LC 96:

Better than MP2 for no items, worse for no items, equivalent for 10 items.
Better than MP3 for 1 item, worse for no items, equivalent for 9 items.

AAC SSR 128:

Better than MP2 for 1 item, worse for no items, equivalent for 9 items.
Better than MP3 for 2 items, worse for no items, equivalent for 9 items.

Thus, we see that only the Main 96 codec is outperformed by any MP2 or MP3 codec for any of these examples. For many programme items, an AAC coder gives statistically superior results. Note that for items Tracy Chapman, Ornette Coleman and Dire Straits there were no significant differences between codecs – all codecs performed the same on these examples.

10.6. Comparison with MPEG-1 codecs

"Is the performance of AAC codecs at the tested bitrate equal to or better than the performance of MPEG-1 Layer II and Layer III?" The accumulated results by codec are shown in Figure 5 (note the foreshortened vertical scale).

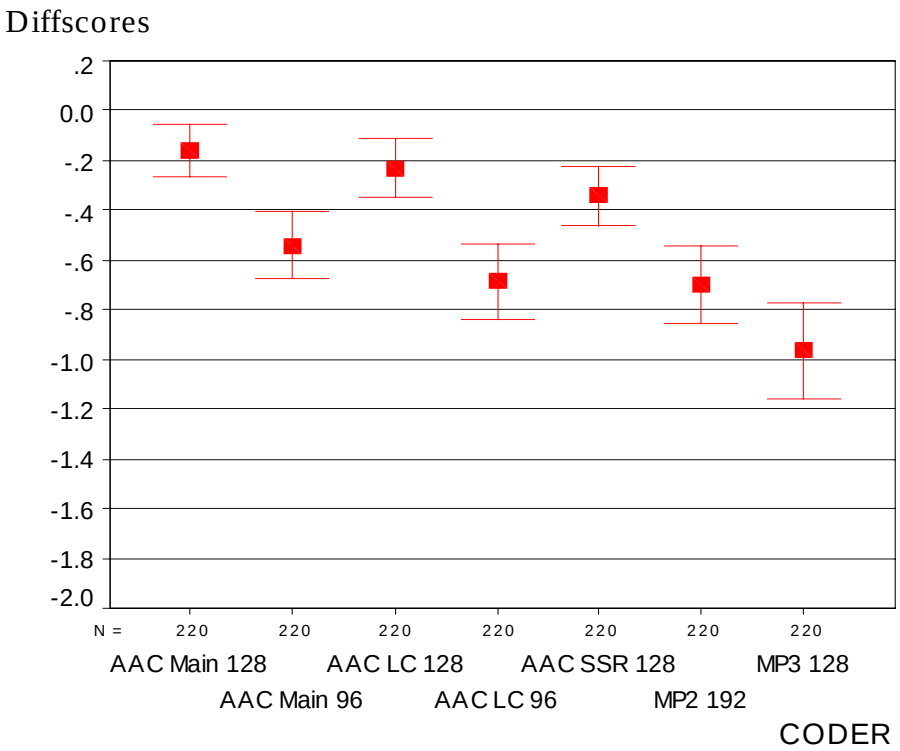


Figure 5. Overall results (averaged across programme items and position) for each coder.

We see from this figure that overall, AAC Main 128, AAC LC 128, and AAC SSR 128 give significantly better performance than do MP2 192 or MP3 128. In addition, AAC Main 96 gives better results than MP3 128. There is no statistically significant improvement between AAC LC 96 and the MPEG-1 codecs.

Within the AAC codec group, AAC Main 128, AAC LC 128, and AAC SSR 128 are all superior to AAC LC 96. In addition, AAC Main 128 and AAC LC 128 are superior to AAC Main 96.

10.7. Statistical indistinguishability

"Is the performance of the coding of AAC codecs at the test bitrate distinguishable from the original signal?" In general, from Figure 5, we see that the performance of the AAC codecs is statistically distinguishable from the original signal. However, for certain items, the codecs give indistinguishable performance. The AAC Main 128 codec is indistinguishable from the original for 8 of 10 items, the AAC Main 96 for 3 items, the AAC LC 128 for 8 of the 10 items, AAC LC 96 for 4 items, and AAC SSR 128 for 8 items. For comparison, MPEG-1 Layer II was statistically indistinguishable for 4 items, and MPEG-1 Layer III for 3 items.

10.8. EBU "Indistinguishable quality"

"Is the performance of AAC codecs at the tested bitrate achieving 'indistinguishable quality' in the EBU definition of that phrase?"

A detailed description of this criterion and the statistical tests required to analyse it are found in [1]. The following chart is in the same format as the chart in section 6.9 of that document. Rather than calculate a "cut-off" as described there, we are directly comparing the confidence intervals as recommended by the EBU definitions. The data for these comparisons is found in Annex 7.

Codec	Items failing	Ratio (if needed)
AAC Main 128	4	0.9528
	9	0.9448
AAC Main 96	0,1,4,8,9	
AAC LC 128	1	0.8931
	2	0.8501
AAC LC 96	0,1,2,3,4,6	
AAC SSR 128	1	0.8436
	2	0.8420
MP2 192	1,2,3,4,7	
MP3 128	0,1,2,6,8,9	

Thus, we see that the AAC Main 128 and AAC LC 128 codecs provide 'indistinguishable quality' in the EBU sense of the phrase[4]. The AAC SSR 128 codec fails to meet this criterion by a margin of less than 1% relative to the decision criterion.

10.9. Most critical material

"Is the following requirement of ITU-R Recommendation BS.1115 fulfilled? 'For emission, the most critical material for the codecs must be such that the degradation may be perceptible but not annoying (grade 4)'"

It is difficult to know exactly what statistical criterion to use in evaluation of this question. The following table shows the mean and lower bound of the confidence interval of the rating score of the most critical (i.e., lowest-rated) test item for each codec, extracted from the table in Annex 7.

[4] The EBU requires 40 subjects to be in the test group. This criterion was not met, and has not been met in any previous test.

Codec	Mean	Lower bound
AAC Main 128	4.4227	4.1051
AAC Main 96	3.4818	3.0077
AAC LC 128	3.8500	3.4409
AAC LC 96	3.1409	2.5833
AAC SSR 128	3.7409	3.2518
MP2 192	2.3182	1.9423
MP3 128	1.6318	1.3105

It is clear that the AAC Main 128 codec meets this criteria, since it is statistically unlikely that for any of the critical items, the true rating is as bad as "perceptible but not annoying". For the cases of AAC LC 96, MP2, and MP3, in each case the confidence interval contains "slightly annoying". The other three cases (AAC Main 96, AAC LC 128, and AAC SSR 128) do not contain "slightly annoying" in the confidence interval, but each has a mean lower than "perceptible but not annoying".

10.10. Ranking of codecs

"What is the relative ranking of the codecs tested?"

A simple comparison of the means gives the following ranking: AAC Main 128, AAC LC 128, AAC SSR 128, AAC Main 96, AAC LC 96, MP2 192, MP3 128. However, there are no statistically significant differences between each pair in this ordering. The gaps indicating statistically-significant differences ($p < 0.05$) occur as follows.

Each coder at or above AAC LC 128 (in the simple ranking above) is better than each coder at or below AAC Main 96.

Each coder at or above AAC SSR 128 is better than each coder at or below AAC LC 96.

Each coder at or above AAC Main 96 is better than each coder at or below MP3 128.

Note, though, that certain items perform differently than is indicated by this one-dimensional ranking. [1] suggests a ranking criterion based on the confidence intervals. Thus the number of items for each coder for which the confidence interval of the diffscore contains 0, and for which it contains a value less than –1 has been tabulated:

Coder	Contains 0	Contains –1
AAC Main 128	7	0
AAC LC 128	7	2
AAC SSR 128	7	2
AAC Main 96	4	4
AAC LC 96	4	3
MP2 192	3	3
MP3 128	3	4

According to this criterion, AAC Main 128 is still clearly the best-judged codec, with AAC LC 128 and AAC SSR included in the next tier. However, the rankings of AAC Main 96, LC 96, MP2, and MP3 are less clear using this system.

11.Test results

The overall test results, averaged across programme items and listener position, are given in Figure 5. The data resulting from the statistical analysis is given in Annex 5, and the graphical presentations, grouped according to programme item, are given in Annex 6.

Here the results are re-presented, grouped according to coder.

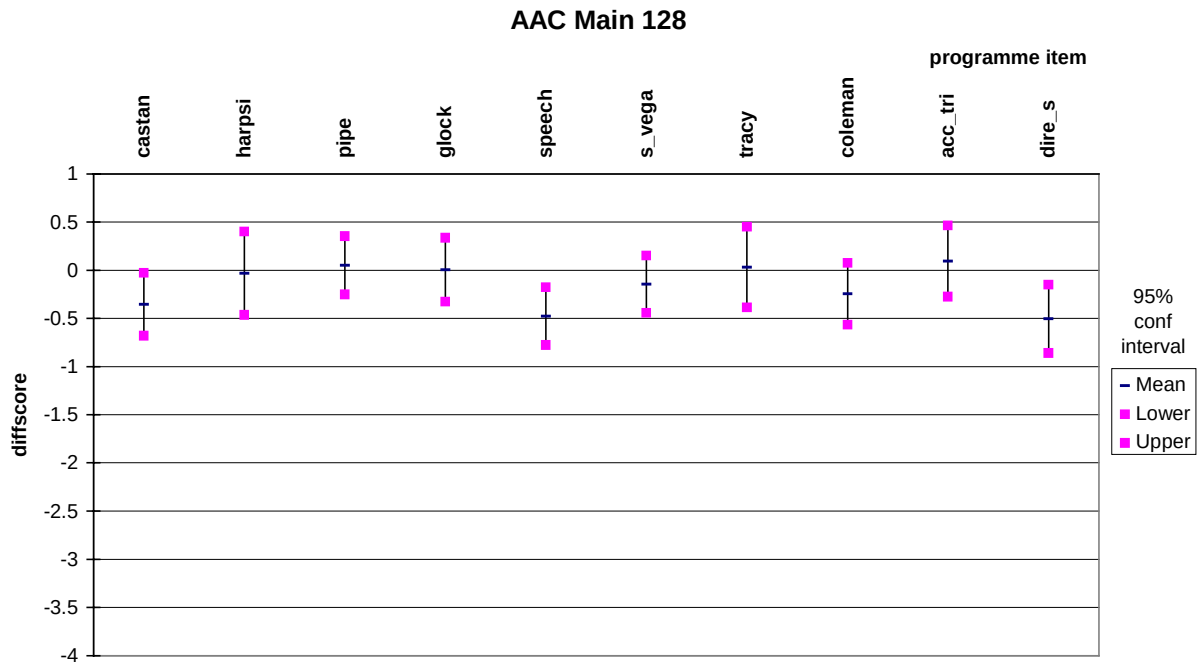


Figure 6. Results for AAC Main Profile at 128 kbps

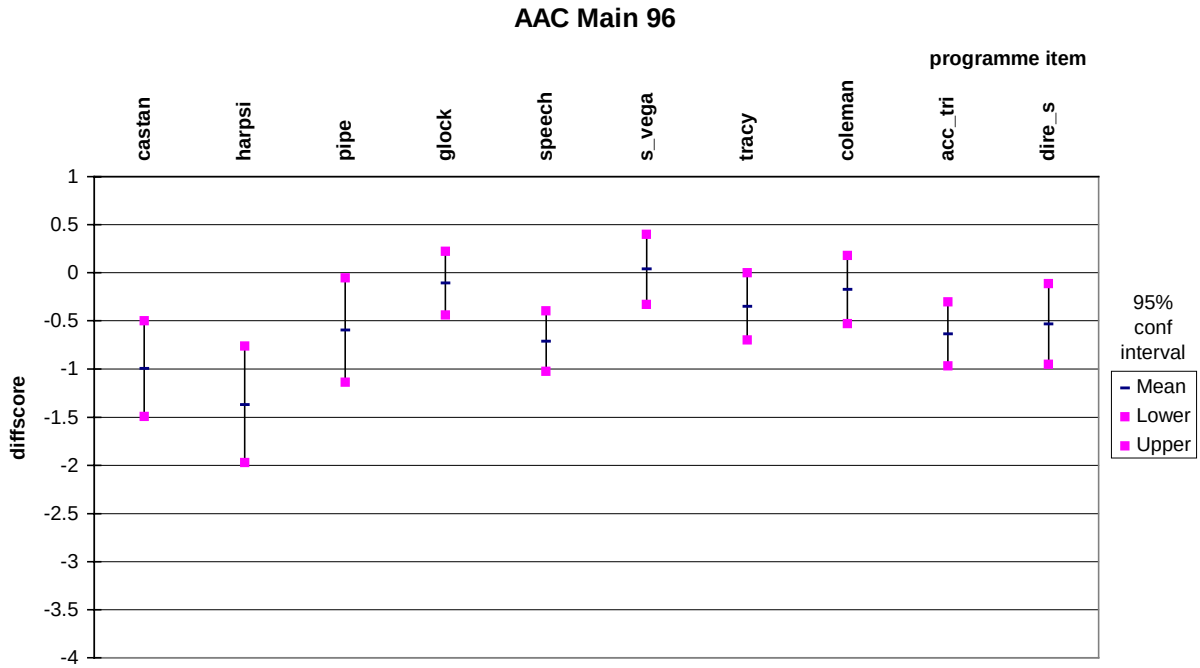


Figure 7 Results for AAC Main Profile at 96 kbps

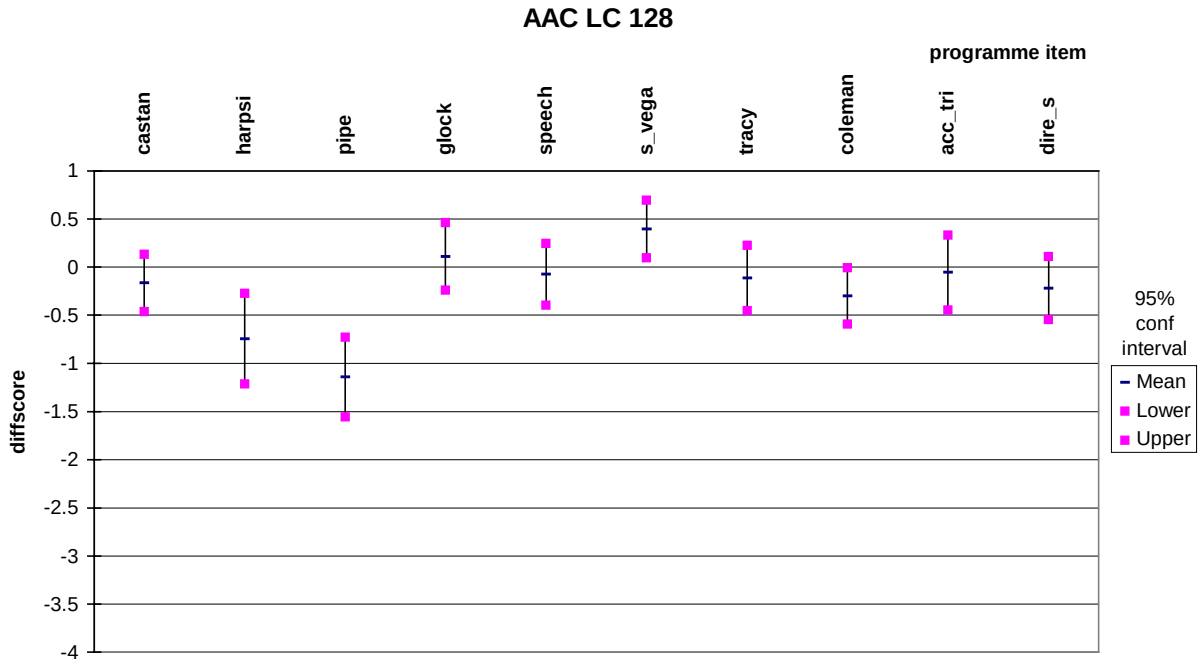


Figure 8 Results for AAC Low Complexity Profile at 128 kbps

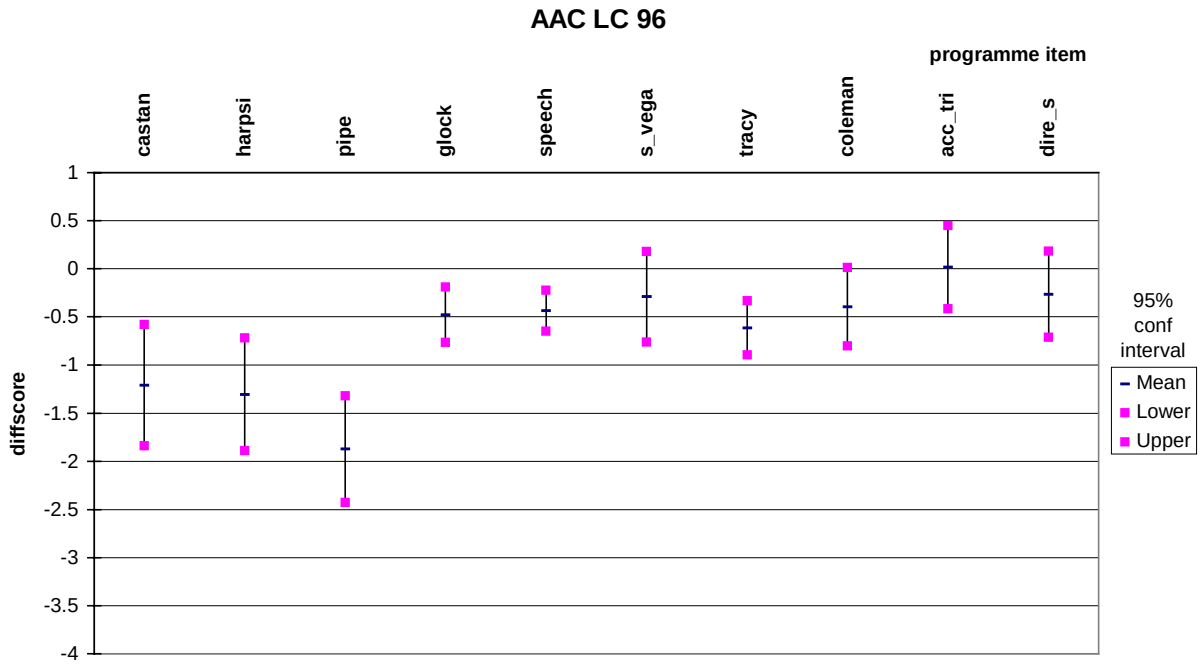


Figure 9 Results for AAC Low Complexity Profile at 96 kbps

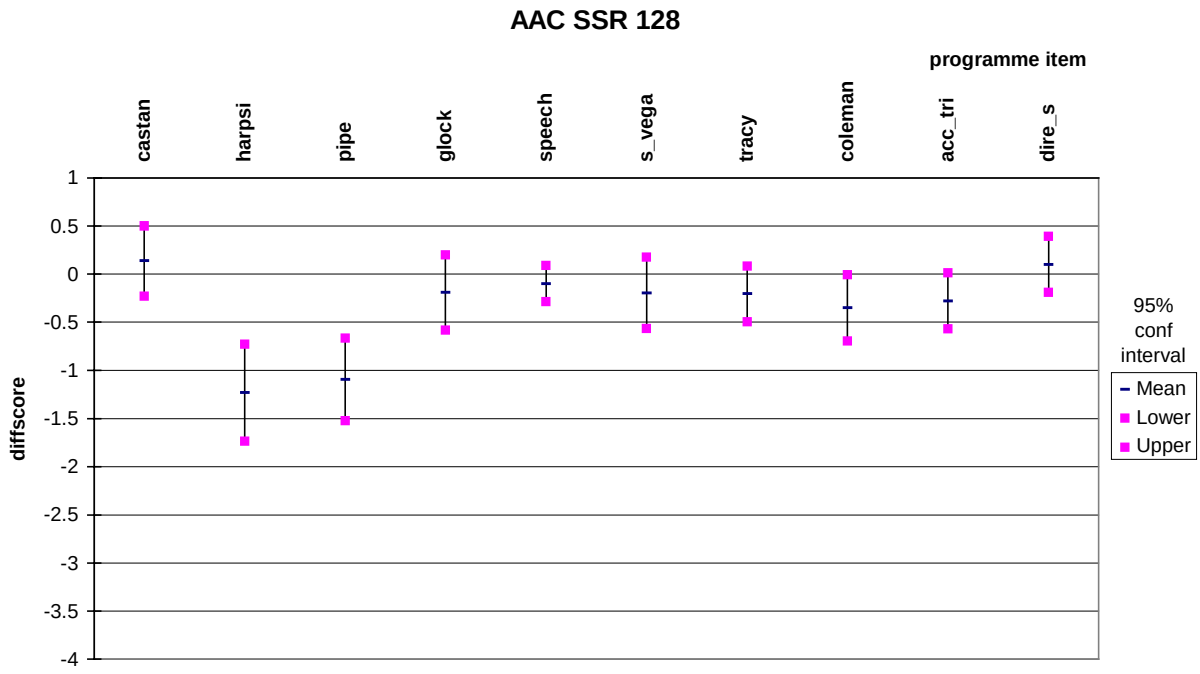


Figure 10 Results for AAC SSR Profile at 128 kbps

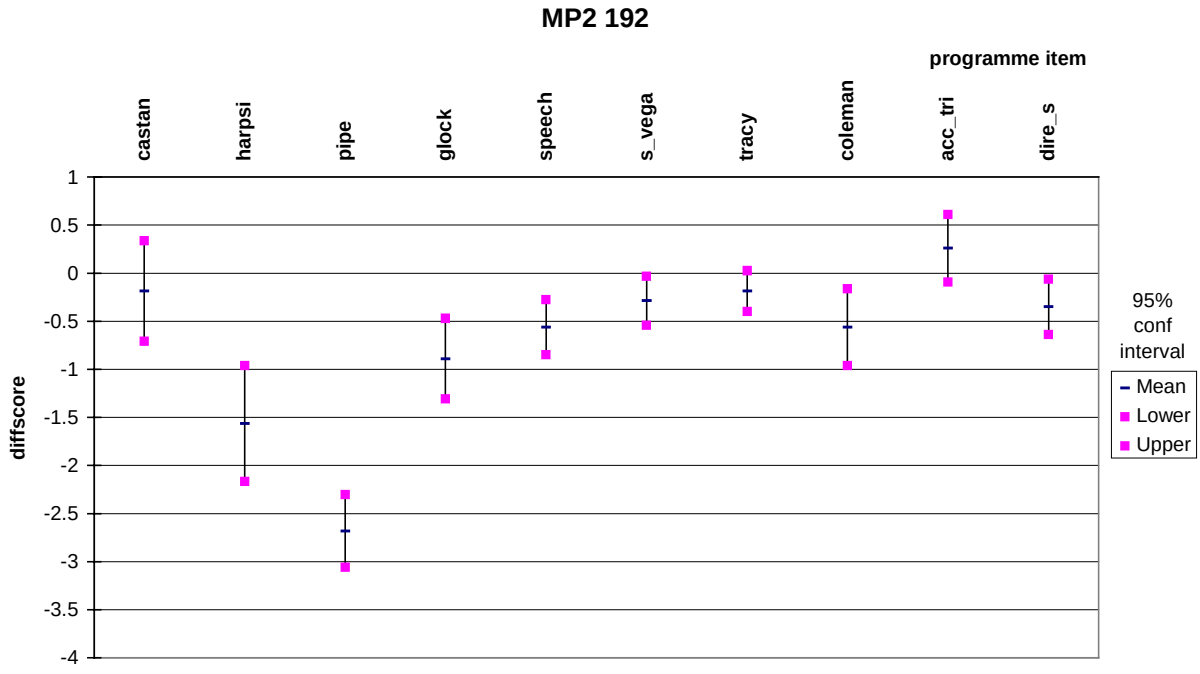


Figure 11 Results for MPEG-1 Layer II at 192 kbps

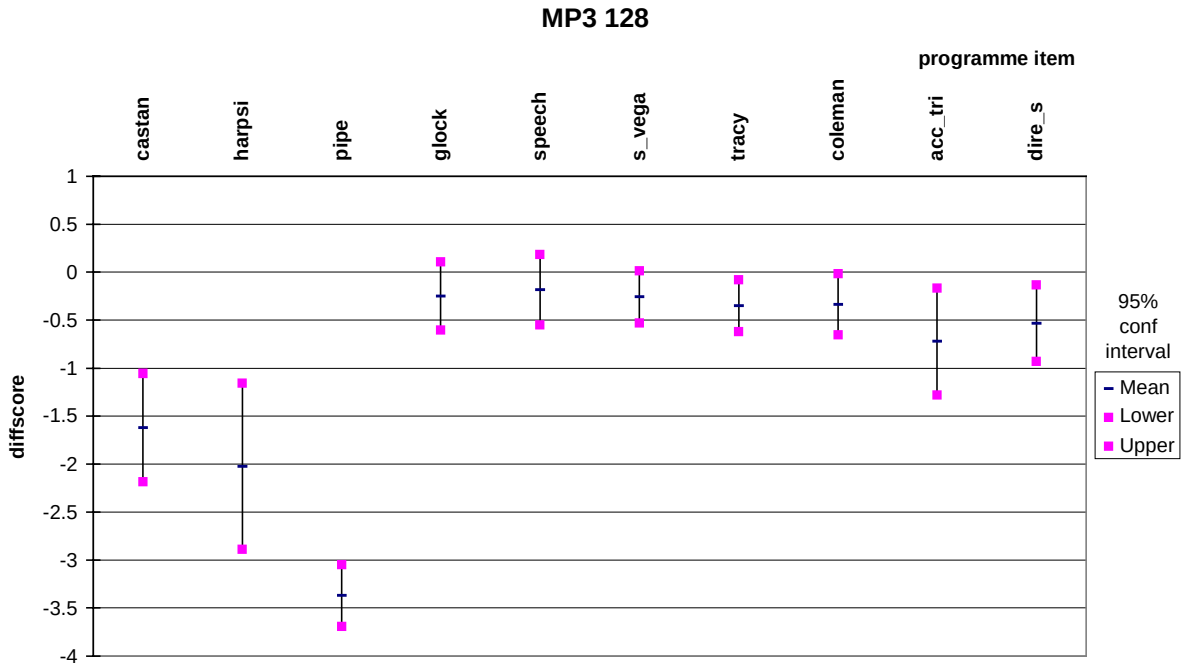


Figure 12 Results for MPEG-1 Layer III at 128 kbps

12.Conclusions

The assessment of the stereo performance of AAC Main, Low Complexity and SSR Profiles have been carried out in comparison with one another and with MPEG-1 codecs at various representative bitrates. The conduct of these tests involved the mutual co-operation and support of a large number of MPEG members and their organisations.

The overall conclusion is that , when auditioning using loudspeakers, AAC coding according to the ISO/IEC 13818-7 standard gives a level of stereo performance superior to that given by MPEG-1 Layer II and Layer III coders.

The test process was intended to answer the following set of questions which form the detailed conclusions of this study:

Is the performance of AAC codecs at the tested bitrate equal to or better than the performance of MPEG-1 Layer II and Layer III?

Section 10.6 presents the answer to this. Overall, all AAC profiles at 128 kbps give significantly better performance than do MPEG-1 Layer II at 192 kbps or Layer III at 128 kbps. Therefore the goal of high audio quality at 64 kbps per channel for MPEG-2 AAC has been achieved. Both AAC Main Profile and Low Complexity Profile provide quality at 96 kbps that is comparable to MPEG-1 Layer II at 192 kbps, and therefore give a 2 to 1 compression advantage. In addition, AAC Main Profile at 96 kbps gives better results than MPEG-1 Layer III at 128 kbps.

Are the listeners' results reliable, i.e. distinguishable from random votes?

The analysis conducted in Section 10.2 concludes that all listeners returned reliable results.

Does the test methodology allow meaningful conclusions to be drawn from these results?

The analysis confirms that meaningful conclusions can be drawn from these results. The effect SEAT was shown to be significant in its own right, particularly with reference to the rearmost seat position. However, as there were only a limited number of listeners who used this position and as the results from the other two seats could be combined, the further analysis only made use of the results from the front and centre seat positions.

How does the performance of the codecs vary with programme items?

This is shown in full in section 11. It is shown that all coders perform, to some extent, differently depending on the type of programme item with which they are being tested.

Is the performance of the coding of AAC codecs at the tested bitrate distinguishable from the original signal?

Sections 10.7 and 11 show that there is a statistical difference between the source and coded items, both overall and for some specific items. However, there were a large number of items for which no difference was recorded.

Is the performance of AAC codecs at the tested bitrate achieving 'indistinguishable quality' in the EBU definition of that phrase?

AAC Main Profile at 128 kbps and AAC Low Complexity Profile at 128 kbps both provided 'indistinguishable quality' and AAC SSR Profile at 128 kbps failed to achieve this by a margin of less than 1% relative to the decision criterion.

Is the following requirement of ITU-R Recommendation BS.1115 [5] fulfilled?
"For emission, the most critical material for the codecs must be such that the degradation may be 'perceptible but not annoying' (grade 4)"

AAC Main Profile at 128 kbps passes this criterion.

What is the relative ranking of the codecs tested?

The relative rankings are presented in section 10.10.

Are there any other features from the data that should be reported?

Comments relating to the significance of listener position are reported in Section 10.3.

13.References

1. Kirby, D. and Watanabe, K., 1996. Report on the formal subjective listening tests of MPEG-2 NBC multichannel audio coding. ISO/IEC JTC1/SC29/WG11/N1419. November 1996.
2. Feige, F., Watanabe, K., Thom, D., Contin, L. 1996. Revised specifications of the MPEG-2 AAC Stereo Verification Tests. ISO/IEC JTC1/SC29/WG11/N1845. October 1997.
3. Audio Subgroup, 1997. Call for new stereo audio test sequences. ISO/IEC JTC1/SC29/WG11/N1706. April 1997
4. ITU-R, 1994. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. ITU-R Recommendation BS.1116. Geneva, 1994.
5. EBU, 1991. 'Basic audio quality requirements for digital audio bit-rate reduction systems for broadcast emission and primary distribution.' CCIR document number TG 10-2/3, 28 October 1991
6. ITU-R, 1994. Low bit-rate audio coding. ITU-R Recommendation BS.1115. Geneva, 1994.

14.Acknowledgements

The authors of this report would like to thank the following additional people, and their companies, for their contributions to the completion of these tests.

Selection panel	A. McParland, T. Buchholz, L. Ferekidis,	BBC R&D Deutsche Telekom, Berkom University of Hannover
	J. Johnston (supervisor)	AT&T
Bitrate verification	M. Coleman S. Quackenbush A. McParland	FiveBats AT&T BBC R&D
Encoder/decoder verification	S. Quackenbush H. Fukuchi F. Feige	AT&T Nippon Steel Deutsche Telekom, Berkom
Preparation of test tapes	S. Quackenbush	AT&T
Conduct of tests	T. Komori	NHK
Test supervisor & mentor	P. Schreiner III	Scientific Atlanta

Annex 1. Report of the Selection Panel for the MPEG-2 AAC Stereo Verification Tests

1.Tasks assigned to the selection panel

1.1. Selection of ten test excerpts

The selection panel will be asked to determine the ten most critical items while avoiding similar material, e.g. bell and triangle should not both be included.

1.2. Selection of training excerpts

Based on past experience, some of the coded test excerpts will be used for the training session. The selection panel will therefore be asked to recommend which of the selected test excerpts should be used for training, bearing in mind that this subset should provoke the range of artefacts likely in the tests and must be fair (i.e. similarly critical) for each codec.

If these criteria cannot be met, then the selection panel should recommend four other critical items for the training session.

For the training the selection excerpts need to ensure that artefacts are clear.

1.3. Selection of low anchor excerpts

As it is anticipated that the results of these tests will show very high quality, it is crucial to prove also that the test was able to reveal artefacts had they been present, otherwise the whole test is invalid. In order to achieve this, we need to include low anchor presentations where the artefacts are a little more obvious, i.e. at about grade 3.5 or so, but not below grade 3.0 (otherwise the grades given to the better presentations will be pushed higher since they appear to be so much better than those below grade 3.0).

A second requirement for including low-anchor stimuli is that we can use these for the assessment of listener reliability (for example using the t-test as suggested in Appendix 1 of BS 1116). This has proved difficult in the past, either because some codec/item combinations were too easy or too difficult. Having some stimuli which are "mid-range" gives grades which can be used more reliably for this assessment.

The selection panel will be asked to ensure that the range of selected codec/item combinations includes a number which is likely to invoke grades in the region of 3 to 3.5. To avoid possible bias during the grading phase of these tests, the identity of low anchor combinations must be kept secret.

1.4. Additional tasks for the selection panel

The selection panel will also be asked to:

- identify any codec/bitrate combinations which consistently offer poor quality and could therefore influence the grading of the remaining codecs. Inclusion of these could be detrimental to the tests and they should therefore be excluded. Presentations which are likely to give grades below 2.5, possibly even 3.0, would be in this category.
- offer advice concerning the tests having auditioned the test excerpts.

2.Conclusions

2.1. Selection of the 10 most critical items

The following 10 items were found to be critical for all of the codecs under test by the selection panel. The details of the selection process are described in the Appendix.

No.	Name	Description
1	Castanets	Castanets
2	Harpsichord	Harpsichord
3	Pitch Pipe	Pitch Pipe
4	Glockenspiel	Glockenspiel
5	Suzanne Vega	Female vocal
6	Male German speech	Male German speech
7	Tracy Chapman	Female voice, Percussion, Synthesiser
8	Ornette Coleman	Saxophone, Trumpet, Double bass, Cymbal
9	Accordion/Triangle	Accordion and Triangle
10	Dire Straits	Synthesiser, High-hat, Drums, Percussion

If a reduction in the number of test signals is necessary, we suggest that the following items could be removed, in this order:

Dire Straits

Suzanne Vega

Male Speech

Ornette Coleman

The listening panel would prefer that no more than two test signals be removed.

2.2. Artefacts observed with the 10 selected critical items

The artefacts for each item are listed roughly in the order in which they were most easily observed. See the Appendix for an explanation of the terms used.

No.	Piece	Artefacts
1	Castanets	Temporal distortion, High frequency loss, High frequency distortion
2	Harpsichord	Temporal distortion, Signal correlated noise, High frequency loss
3	Pitch Pipe	Distortion, Signal correlated noise, High frequency loss, Periodic modulation
4	Glockenspiel	Temporal distortion, High frequency loss, Signal correlated noise
5	Male German speech	Signal correlated noise, High frequency loss
6	Suzanne Vega	High frequency loss, High frequency excess (sibilance), Signal correlated noise
7	Tracy Chapman	Signal correlated noise, Image quality, High frequency loss
8	Ornette Coleman	Image quality, Periodic modulation, High frequency loss, Signal correlated noise
9	Accordion/Triangle	Image quality, Signal correlated noise, Distortion, High frequency loss
10	Dire Straits	Image quality, High frequency distortion, High frequency loss

Artefact categories for each codec

This table contains a list of the main artefacts found in each codec for each item. The artefacts are listed in approximate order of severity. See the Appendix for the numbers corresponding to the artefact categories.

Item/Codec	A	B	C	D	E	F	G	H
Castanets	5, 9, 2	5, 4, 2	5, 3, 9	5, 9	5, 2, 9	5, 2	5, 2	5
Harpsichord	5, 4	5, 2	3, 1	4, 1, 5	2, 5	5, 2 ,1	5, 2 ,1	2, 1
Pitch Pipe	4, 2, 6, 1	2, 6	3, 1	4, 2, 6, 1	2, 6, 1	4, 1	1, 2	1
Glockenspiel	5, 2, 1	5, 2	5, 3, 1	5, 2, 1	5, 2	5, 1	1, 2, 5	5, 1
German male speech	1, 2, 9, 4	1, 2	1, 3	1, 2	1, 2	2, 3, 1	1, 5, 2	2, 1
Suzanne Vega	2, 3, 1	2, 3, 1	3, 1	2, 3, 1	2, 3, 1	3, 1	2	1, 2
Tracy Chapman	1, 8, 2	1, 2	1, 3, 8	1, 8, 2	1, 2	2, 1	2, 1	9, 1
Ornette Coleman	8, 4, 2, 7	8, 2	8, 3, 1	8, 4, 2, 1	4, 2, 8	1, 8	4, 1	4, 8
Accordion/Triangle	8, 6, 1, 2	2, 6, 1	3, 8, 1	8, 6, 1, 2	2, 1, 6	1, 8	2, 1, 8	2, 8, 1
Dire Straits	9, 8	2, 8	8, 3	8, 9	2, 8, 9	8, 9	8, 2	1, 8

Summary of main characteristics

Codec A: Dominated by signal correlated noise with loss of high frequency and poor image quality, but also periodic modulation effects.

Codec B: Mainly loss of high frequency with signal correlated noise and temporal distortions.

Codec C: Dominated by excess of high frequency followed by signal correlated noise and image quality.

Codec D: Dominated by signal correlated noise with loss of high frequency, poor image quality and periodic modulation effects, but also temporal distortions.

Codec E: Dominated by loss of high frequency with signal correlated noise and temporal distortions.

Codec F: Dominated by signal correlated noise with loss of high frequency, but also poor image quality and temporal distortions.

Codec G: Dominated by signal correlated noise with loss of high frequency, but also temporal distortions.

Codec H: Dominated by signal correlated noise with loss of high frequency, but also poor image.

2.3. Training Items

The following four of the selected ten most critical items are recommended for training of the test subjects.

No.	Name
1	Tracy Chapman
2	Ornette Coleman
3	Castanets
4	Pitch Pipe

If time and resources permit, training listeners on the accordion/triangle signal would also be advantageous, because it is a new signal, and the distortions are somewhat difficult to hear without familiarity to the original. If Ornette Coleman is removed as a test item, we suggest that the accordion/triangle signal replace it.

2.4. Low anchors

After listening to the signals at both high and low bitrates, we feel that significant low anchors are already included in the data.

2.5. Poor quality codecs

The selection panel rejected one bitrate/codec combination. This codec should not be included in the test, as it will introduce a very low (1-2) anchor. The panel was not informed of the identity of the rejected codec until after the selection phase and rejection decision were completed.

2.6. Advice concerning the test

The selection panel and the administrator both feel that the quality of some of the codecs is good enough that a switched ABC hidden-reference method is more appropriate than the sequential ABC hidden-reference method. For some of the codecs, sequential presentation will reduce the test sensitivity.

The selection panel has also listened to the selected critical items with speakers as well as headphones, and notes that the loudspeaker presentation exposes distortions and imaging problems that are not easily heard in the headphone test. If at all possible, a loudspeaker test is encouraged. The listening panel feels that scores for some items will be lower if the test is performed using loudspeakers. One listener observed that artefacts were more audible if he moved toward the back of the room, perhaps beyond the critical distance in the listening room.

The administrator and selection panel would like to thank Fraunhofer IIS, Joachim Gnauk, Martin Dietz, and Olivier Kunz for their support and hospitality.

During the initial pre-screening phase, it was noted that one codec, identified by the administrator (after its performance was questioned) as the Low Complexity Profile, appeared to be broken. As Mr. Quackenbush of AT&T was not available to replace the encoded material, the LC profile material was replaced immediately by FhG, with the agreement of Mr. Johnston of AT&T. This coder was replaced before the screening process was started, as the distortions in the supplied material were substantial, and unlike normal coding artefacts. The decision on test material, anchors, and the like were made with the FhG low complexity profile codec. The identity of the replaced codec was obscured from the selection panel.

3.Appendix: Details of the Selection Process

3.1. Listening room and technical equipment

The listening room and test equipment were provided by FhG, who will provide a description if requested. The listening panel felt that the equipment and situation were quite sufficient for the pre-screening task.

3.2. Item list reduction process

The listening panel and administrator listened to all of the 42 test signals with each coding system operating at the lower bit rates. The most impaired were noted, and then the panel continued by listening to the higher rate codecs for each of the selected signals. At this point, 22 signals remained in the list. After listening to the high-rate codecs, the 10 most sensitive signals were selected, considering both the low rate and high rate codec performance. This process is tabulated in the “List for Selection of Test Excerpts“ (Section 3.4 of this Appendix).

A PERL script running in a terminal window connected to the SGI workstation enabled the listeners to play any item with a particular set of codecs (initially the low bit-rate versions). It was agreed that each codec would be associated with a particular letter throughout the listening, to aid in summarising the

codec's artefacts. However, it was also thought that a random order of presentation of codecs would help isolate the individual codec characteristics. The initial sequence was Ref, A, B, C, Ref, D, E, with the order of the codecs (A, B, etc.) randomised. Initially, a codec F was included, but that codec was removed due to its quality. For the high bit-rates another script provided Ref, F, Ref, G, Ref, H, with the order of codecs again randomised.

3.3. Impairment Categories Table

This table is derived, with changes, from ISO/IEC JTC1/SC29/WG11 No 685, March 1994.

No.	Artefact Category	Explanation
1	Signal correlated noise	coloured noise associated with the signal
2	Loss of High Frequency	lack of high frequencies
3	Excess of High Frequency	excess of high frequencies or associated effects, e.g. sibilance
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 non-harmonic distortion
7	Extra Sounds	spurious sounds not related to the material, e.g. clicks
8	Image Quality	all aspects including spreading, movement, stability and phase related effects
9	High frequency distortion	phasey distortions in the high frequencies

3.4. List for Selection of Test Excerpts

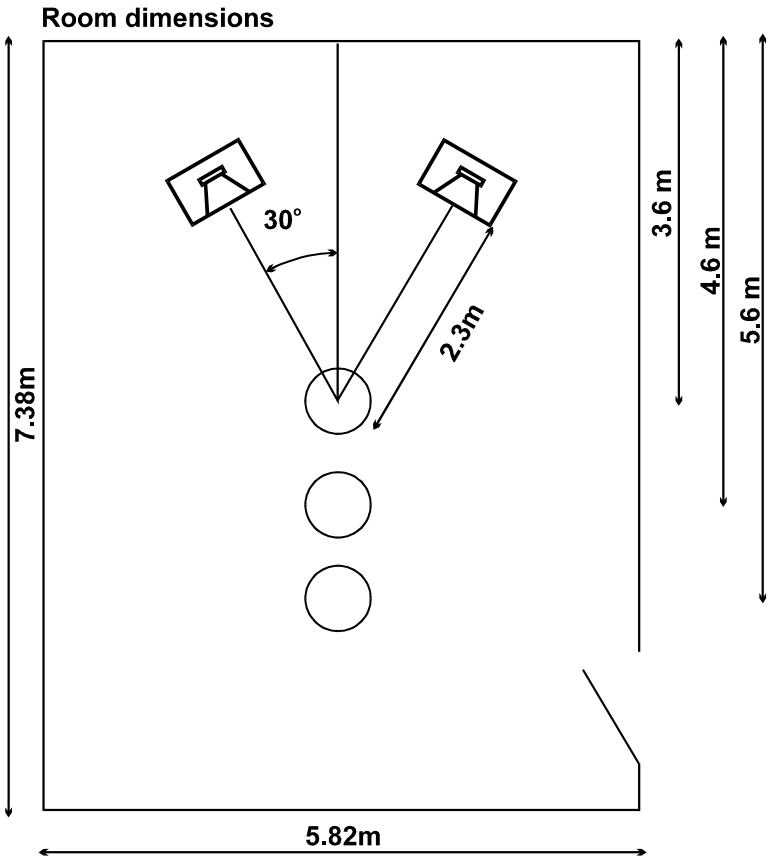
filename	signal	source	1st Step	2nd Step	3rd Step
te1	Dorita	Lou Reed (Magic and Loss)			
te2	We shall be happy	Ry Cooder (Jazz)	X	X	
te3	Castanets	SQAM	X	X	X
te4	Harpsichord	SQAM	X	X	X
te5	Pitch Pipe	Dolby	X	X	X
te6	Glockenspiel	SQAM	X	X	X
te7	Male German Speech	SQAM	X	X	X
te8	Suzanne Vega	Suzanne Vega Toms Diner (Solitude Standing)	X	X	X
te9	Tracy Chapman	Elektra 960 774-2	X	X	X
te10	Fireworks	Pierre Verany 788031	X		
te11	Ornette Coleman	Dreams 008	X	X	X
te12	Bass Synth	RR recording (DAT)			
te13	Bass guitar	RR recording (DAT)			
te14	Hyden Trumpet Concert	Philips 420 203-2			
te15	Carmen	Telarc CD-80224	X		
te16	Accordion/Triangle	Private (analogue) recording	X	X	X
te17	Tambourine	RR recording (DAT)			
te18	Percussion	Sonic Images SICD2026	X		
te19	Male speech	Japan Audio Society CD-3			
te20	George Duke	Elektra 960 778-2	X		
te21	Asa Jinder	Eagle Records, ECD 015	X		
te22	Dire Straits	Warner Bros. 7599-25264-2	X	X	X
te23	Dalarnas Spelmansforbund	Mono Music AB MMCD 005			
te24	Stefan Nilsson	Swedish Radio/Pioneer PIECD-01	X		
te25	Stravinsky	Sony Classical SK45965	X		
te26	Ravel	Telarc CD-80171			
te27	Triangles	SQAM	X	X	
te28	Clay	NTT (Ms. Maiko Iuchi)[5]			

[5] The named person for each item contributed through NTT is the composer.

filename	signal	source	1st Step	2nd Step	3rd Step
te29	spiral wave	NTT (Mr. Shintaro Imai)			
te30	"aimai"	NTT (Ms. Ayako Kashide)	X		
te31	ether	NTT (Mr. Shu Matsuda)			
te32	Palmtop boogie	NTT (Mr. Ken'ichi Sakakibara)	X		
te33	<CROISEMENT I> pour haubois, violon et contrebasse	NTT (Ms. Hitomi Kaneko)			
te34	drifting	NTT (Mr. Naoki Ono)			
te35	dramatics	NTT (Ms. Yoshiko Ando)			
te36	O1	NTT (Ms. Tomoko Nakai)			
te37	Fourth	NTT (Ms. Yuka Yamashita)			
te38	Interlude by Halves for violin, flute and piano	NTT (Ms. Mitsuyo Hashida)			
te39	accellation	NTT (Ms. Chiaki Mouri)			
te40	atmosphere	NTT (Mr. Naotoshi Osaka)			
te41	fanfare	NTT (Mr. Naotoshi Osaka)			
te42	Kids Drive Dance(KDD)	NTT (Mr. George Aburai)	X		

Annex 2. Listening room conditions and equipment

1.Listening Room Conditions



Length:	7.38 m
Width:	5.82 m
Height:	3.30 m
Floor area:	42.95 m^2
Volume:	141.74 m^3
Listening distance:	2.30 m
	3.20 m
	4.15 m
Height of the acoustic centre of the loudspeaker:	1.32 m

Listening Room B268, NHK Science & Technical Research Laboratories.

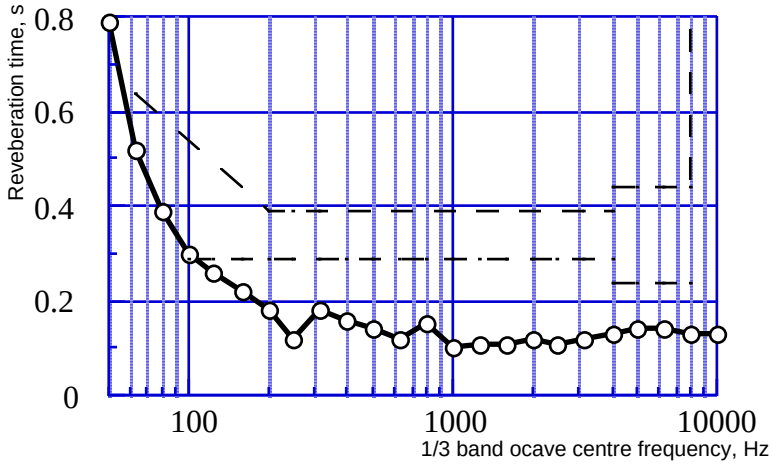
2. Listening Level of the seat positions

seat position	Listening Level dB (A)
1	84.9
2	82.8
3	81.5

3. List of Test Equipment

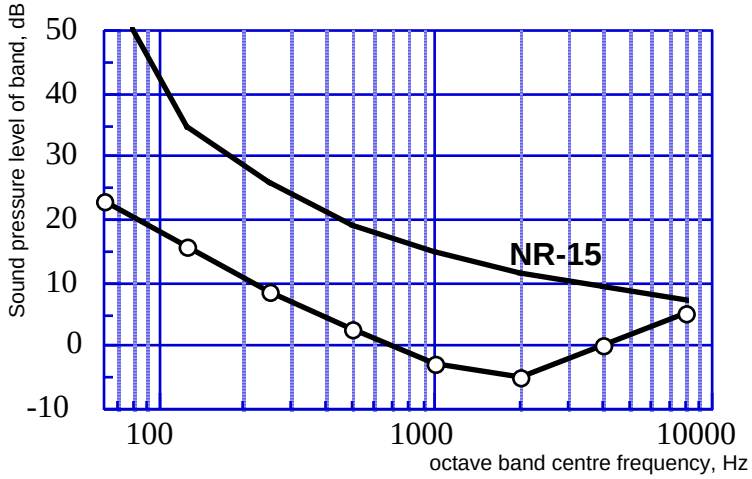
Qty	Description	Model
1	Digital Audio Tape Recorder	SONY PCM-7050
1	D/A Converter Unit	DCS 952
2	Loudspeaker Unit	Mitsubishi 2S-3003
1	Amplifier	Accuphase PRO-20

4.Reverberation time



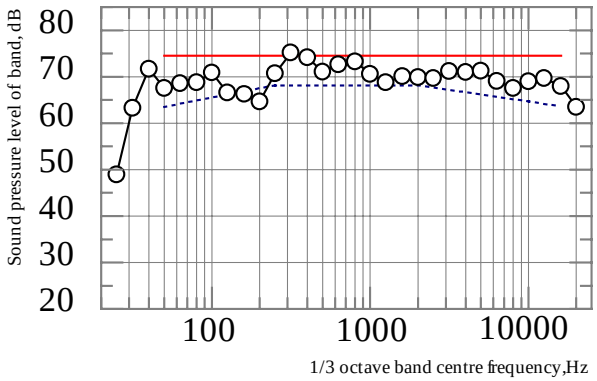
The mean reverberation time between 200 Hz and 4 kHz is 0.13s, which is below the range recommended in BS-1116 for this size of room.

5.Background noise

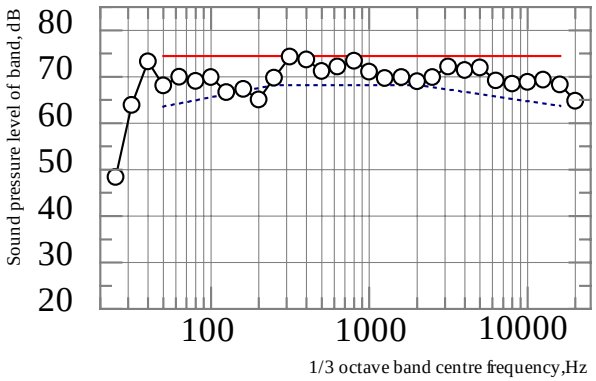


The noise level at the reference listening position meets the noise criterion NR-15.

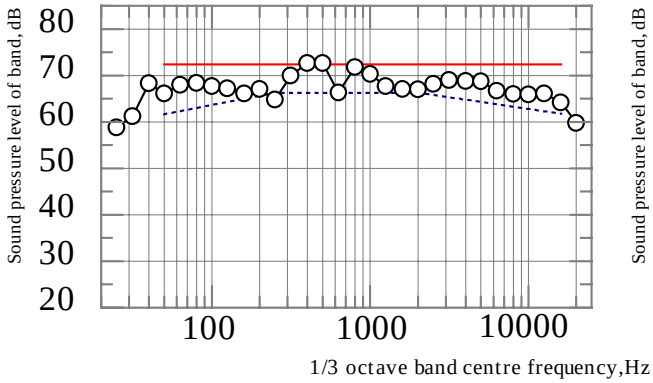
6.Frequency response measurements



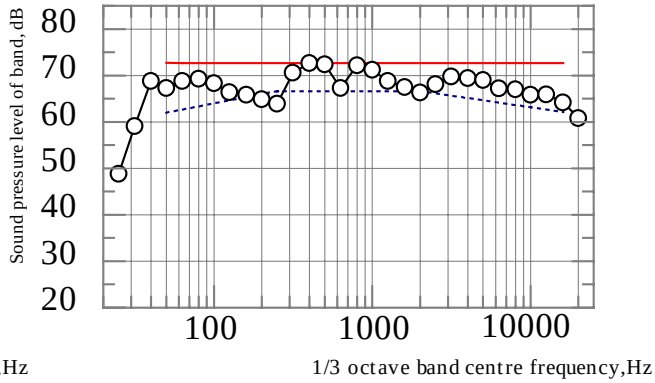
1st position Left channel



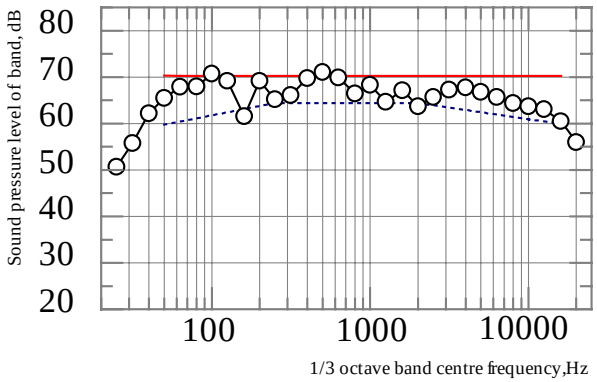
1st position Right channel



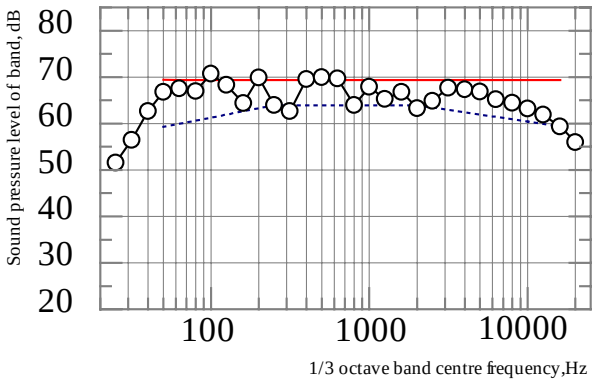
2nd position Left channel



2nd position Right channel



3rd position Left channel



3rd position Right channel

Annex 3. List of Participants

Date	Surame	First name	Organization	Experience of listening test	Age	Male/ Female	Seat Position
97.11.20-21	Ohta	Yasuji	Fujitsu Laboratories Limited	Yes	33	M	1
97.11.20-21	Ohara	Hiroyuki	Ricoh Company,LTD	No	27	M	3
97.11.20-21	Araki	Tadashi	Ricoh Company,LTD	No	33	M	2
97.11.20-21	Saito	Takashi	NHK Broadcasting Engineering Dept.	Yes	21	M	1
97.11.20-21	Ono	Kazuho	NHK Science & Technical Research Labs.	Yes	32	M	2
97.11.20-21	Masaoka	Kenichiro	NHK Science & Technical Research Labs.	Yes	25	M	1
97.11.20-21	Chiba	Shinichi	NHK Science & Technical Research Labs.	Yes	29	M	2
97.11.25-26	Emi	Tetsuro	Pioneer Electronic Corporation	No	38	M	1
97.11.25-26	Fujiyoshi	Akimitsu	Pioneer Electronic Corporation	No	30	M	3
97.11.25-26	Suzuki	Masami	Pioneer Electronic Corporation	Yes	35	M	2
97.11.27-28	Wada	Masuo	Toshiba Corporation	Yes	52	M	2
97.11.27-28	Kaneko	Itaru	ASCII Corporation	Yes	40	M	1
97.11.27-28	Obata	Shinichi	Hitachi,LTD	Yes	30	M	3
97.11.27-28	Yasura	Sadahiro	Victor Company of Japan,Limited	Yes	31	M	2
97.11.27-28	Kuran	Takehiko	Victor Company of Japan,Limited	Yes	30	M	3
97.11.27-28	Nishimoto	Kengo	NHK Broadcasting Engineering Dept.	Yes	35	M	1
97.12.1-2	Takahashi	Fumiko	music academy student	No	19	F	2
97.12.1-2	Sato	Kotono	music academy student	No	19	F	3
97.12.1-2	Hatta	Akiyo	music academy student	No	21	F	1
97.12.1-2	Matsunaga	Eiichi	Fuji Television Network INC	No		M	1
97.12.1-2	Fukumori	Takaharu	Tokyo FM Broadcasting Co.LTD	No	39	M	3
97.12.1-2	Kawabuchi	Tsuyoshi	Tokyo FM Broadcasting Co.LTD	No		M	2
97.12.1-2	Takanose	Hajime	music academy student	No	19	M	2
97.12.1-2	Nishida	Fumiaki	Fujitsu Limited	Yes	29	M	1
97.12.1-2	Tsuboi	Mitsuru	Fujitsu Limited	Yes	32	M	3
97.12.3-4	Hashimoto	Kenichi	Tokyo Broadcasting System, Inc.	Yes	38	M	3
97.12.3-4	Matsuoka	Takeo	Tokyo Broadcasting System, Inc.	No	30	M	2
97.12.3-4	Azuma	Mitsuyoshi	Nippon Television Network Corporation	Yes	41	M	1
97.12.3-4	Katayama	Takashi	Matsushita Electric Industrial Co., Ltd.	Yes		M	2
97.12.3-4	Fujita	Takeshi	Matsushita Electric Industrial Co., Ltd.	Yes		M	1
97.12.3-4	Otani	Masamichi	NHK Science & Technical Research Labs.	Yes	28	M	3

Annex 4: PERL script

The PERL script used to transform the data from the subject-by-item matrix to the flat case listing.

```
#!/usr/bin/perl

$ct = 0;
while (<>) {
  if ($ct < 80) {
    ($trial[$ct], $item[$ct], $coder[$ct], $order[$ct]) = split;
  }
  elsif ($ct == 81) { # seat number
    @seat = split;
    }
  elsif ($ct > 82) {
    $thisitem = $ct - 83;
    @data = split;
    if ($data[0] != $thisitem+1) {
      die "data error.\n";
      }
    for ($i=0;$i!=31;$i++) {
      if ($order[$thisitem] == 1) { # ref first
        $ref = $data[$i*2+1];
        $test = $data[$i*2+2];
        }
      else {
        $ref = $data[$i*2+2];
        $test = $data[$i*2+1];
        }
      printf("%d %d %d %d %.2f %.2f\n",$i,$seat[$i+1],$item[$thisitem],
           $coder[$thisitem],$ref,$test);
    }
  }
  $ct++;
}
```

Annex 5. Means and 95% Confidence Intervals

Item-by-item and coder-by-coder breakdowns of the mean and confidence intervals of the diffscores.

For each, a * indicates that that coder, for that item, gave results which were statistically identical to the test signal (that is, the confidence interval contains 0); a – indicates that that coder, for that item, gave results which might be worse than "perceptible but not annoying" (that is, the confidence interval contains values less than –1).

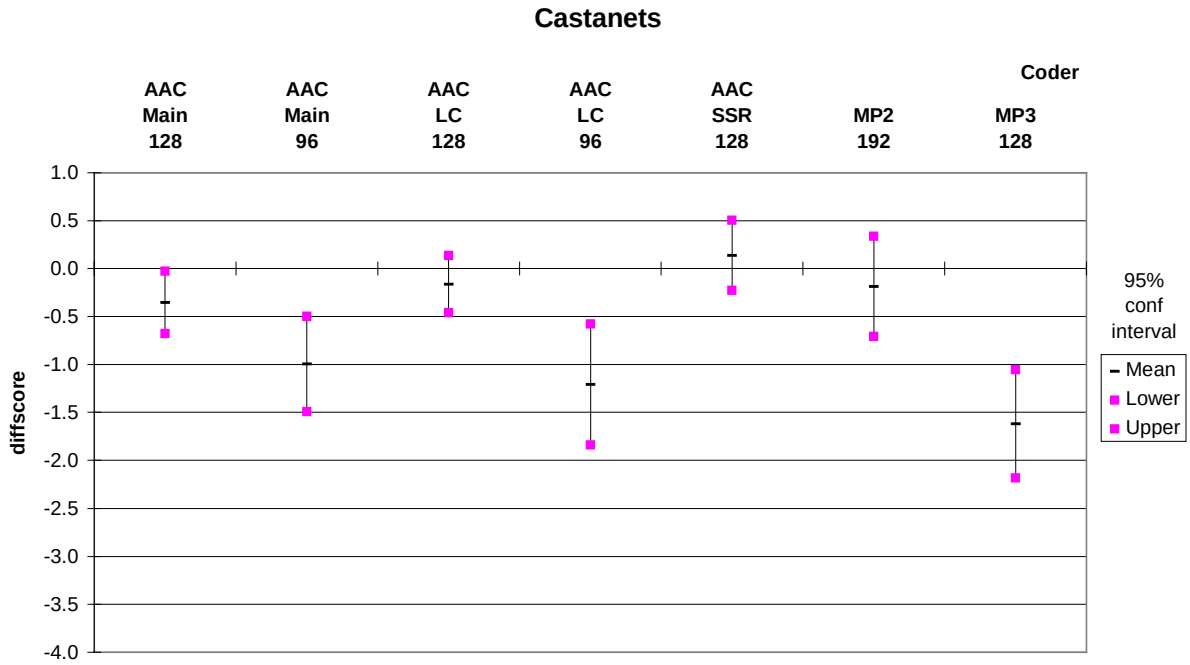
These results are shown graphically in Annex 6.

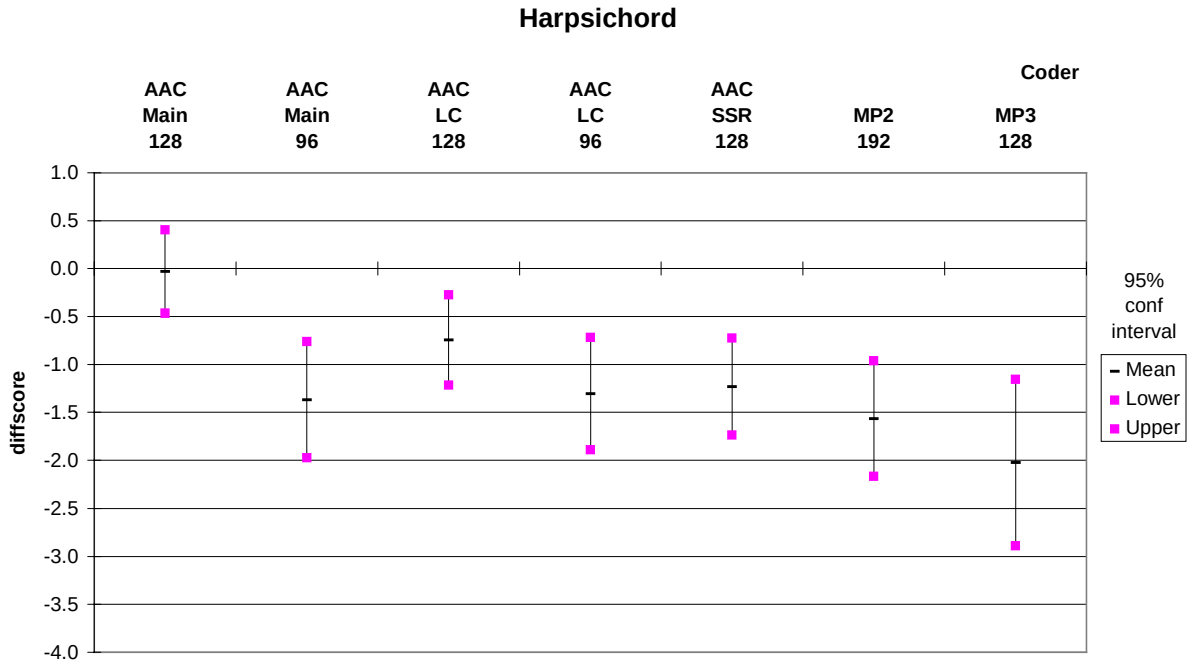
ITEM	CODER	Mean	Confidence interval	
			Lower	Upper
0 - Castanets	AAC Main 128	-0.3545	-0.6809	-0.0282
	- AAC Main 96	-0.9955	-1.4925	-0.4984
	* AAC LC 128	-0.1636	-0.4623	0.1351
	- AAC LC 96	-1.2091	-1.8400	-0.5782
	* AAC SSR 128	0.1364	-0.2296	0.5024
	* MP2 192	-0.1864	-0.7098	0.3370
	- MP3 128	-1.6182	-2.1818	-1.0546
	- codec_x	-0.8727	-1.6342	-0.1112
1 - Harpsichord	* AAC Main 128	-0.0318	-0.4646	0.4010
	- AAC Main 96	-1.3682	-1.9742	-0.7622
	- AAC LC 128	-0.7455	-1.2150	-0.2759
	- AAC LC 96	-1.3045	-1.8912	-0.7179
	- AAC SSR 128	-1.2318	-1.7369	-0.7268
	- MP2 192	-1.5636	-2.1655	-0.9618
	- MP3 128	-2.0227	-2.8900	-1.1554
	- codec_x	-2.8591	-3.3434	-2.3748
2 - Pitch Pipe	* AAC Main 128	0.0500	-0.2522	0.3522
	- AAC Main 96	-0.5955	-1.1361	-0.0548
	- AAC LC 128	-1.1409	-1.5556	-0.7262
	- AAC LC 96	-1.8727	-2.4267	-1.3188
	- AAC SSR 128	-1.0955	-1.5243	-0.6666
	- MP2 192	-2.6818	-3.0577	-2.3059
	- MP3 128	-3.3682	-3.6895	-3.0469
	- codec_x	-3.0545	-3.4940	-2.6150
3 - Glockenspiel	* AAC Main 128	0.0045	-0.3274	0.3365
	* AAC Main 96	-0.1091	-0.4392	0.2210
	* AAC LC 128	0.1091	-0.2427	0.4609
	AAC LC 96	-0.4773	-0.7656	-0.1890
	* AAC SSR 128	-0.1909	-0.5827	0.2009
	- MP2 192	-0.8909	-1.3097	-0.4721
	* MP3 128	-0.2500	-0.6046	0.1046
	- codec_x	-0.6545	-1.1728	-0.1363
4 - Male German Speech	AAC Main 128	-0.4773	-0.7764	-0.1781
	- AAC Main 96	-0.7091	-1.0223	-0.3959
	* AAC LC 128	-0.0727	-0.3928	0.2473
	AAC LC 96	-0.4364	-0.6477	-0.2250
	AAC SSR 128	-0.1000	-0.2871	0.0871
	MP2 192	-0.5636	-0.8515	-0.2758
	* MP3 128	-0.1818	-0.5494	0.1857
	- codec_x	-0.6318	-1.0442	-0.2194
5 - Suzanne Vega	* AAC Main 128	-0.1455	-0.4426	0.1517
	* AAC Main 96	0.0364	-0.3294	0.4021
	AAC LC 128	0.3955	0.0958	0.6951
	* AAC LC 96	-0.2909	-0.7615	0.1797
	* AAC SSR 128	-0.1955	-0.5670	0.1761

ITEM	CODER	Mean	Confidence interval	
			Lower	Upper
	MP2 192	-0.2864	-0.5406	-0.0321
	* MP3 128	-0.2591	-0.5309	0.0127
	* codec_x	-0.3500	-0.7122	0.0122
6 - Tracy Chapman	* AAC Main 128	0.0318	-0.3864	0.4501
	* AAC Main 96	-0.3500	-0.7004	0.0004
	* AAC LC 128	-0.1136	-0.4536	0.2263
	AAC LC 96	-0.6136	-0.8935	-0.3338
	* AAC SSR 128	-0.2045	-0.4930	0.0839
	* MP2 192	-0.1864	-0.3985	0.0257
	MP3 128	-0.3500	-0.6225	-0.0775
	codec_x	-0.4227	-0.7068	-0.1387
7 - Ornette Coleman	* AAC Main 128	-0.2455	-0.5666	0.0757
	* AAC Main 96	-0.1727	-0.5264	0.1810
	AAC LC 128	-0.3000	-0.5906	-0.0094
	* AAC LC 96	-0.3955	-0.8031	0.0122
	AAC SSR 128	-0.3500	-0.6936	-0.0064
	MP2 192	-0.5636	-0.9629	-0.1644
	MP3 128	-0.3364	-0.6560	-0.0168
	codec_x	-0.4955	-0.9140	-0.0770
8 - Accordion & Triangle	* AAC Main 128	0.0955	-0.2754	0.4663
	AAC Main 96	-0.6364	-0.9700	-0.3027
	* AAC LC 128	-0.0545	-0.4449	0.3358
	* AAC LC 96	0.0182	-0.4138	0.4501
	* AAC SSR 128	-0.2773	-0.5698	0.0152
	* MP2 192	0.2591	-0.0902	0.6084
	- MP3 128	-0.7227	-1.2799	-0.1656
	* codec_x	-0.2409	-0.7393	0.2575
9 - Dire Straits	AAC Main 128	-0.5045	-0.8599	-0.1492
	AAC Main 96	-0.5318	-0.9530	-0.1107
	* AAC LC 128	-0.2182	-0.5456	0.1092
	* AAC LC 96	-0.2636	-0.7105	0.1832
	* AAC SSR 128	0.1000	-0.1906	0.3906
	MP2 192	-0.3500	-0.6379	-0.0621
	MP3 128	-0.5318	-0.9308	-0.1328
	codec_x	-0.5591	-0.9674	-0.1507

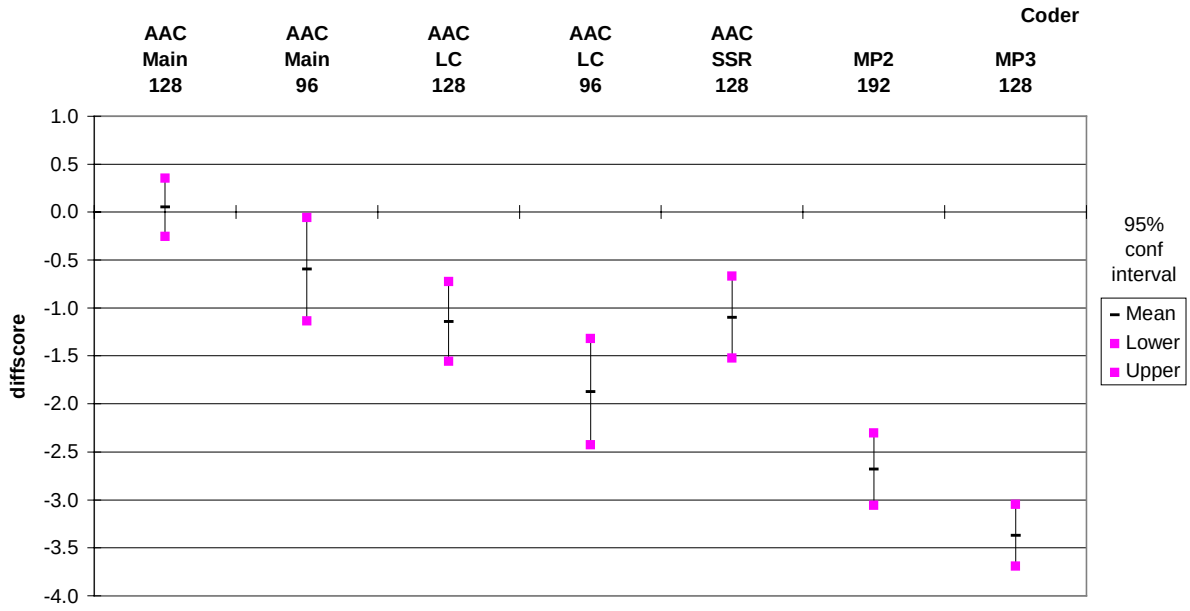
Annex 6 Graphical presentation based on programme item

Graphical presentation of the data in Annex 5. Each plot shows the results of each coder applied to one critical programme item. Error bars indicate 95% confidence intervals. Each data point corresponds to judgements by 22 subjects.

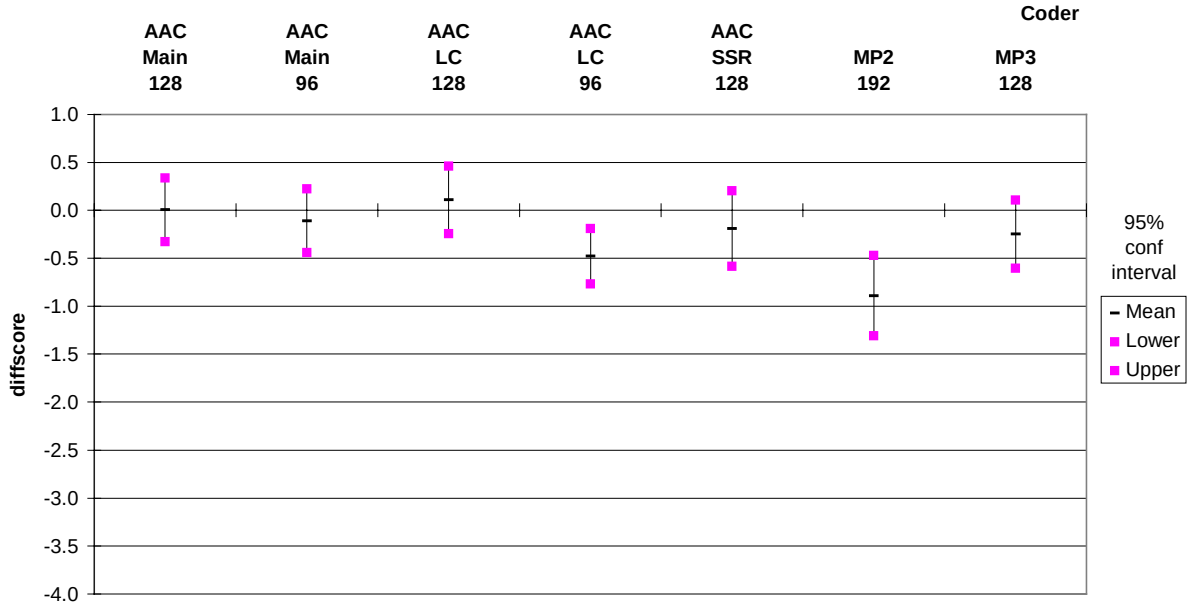




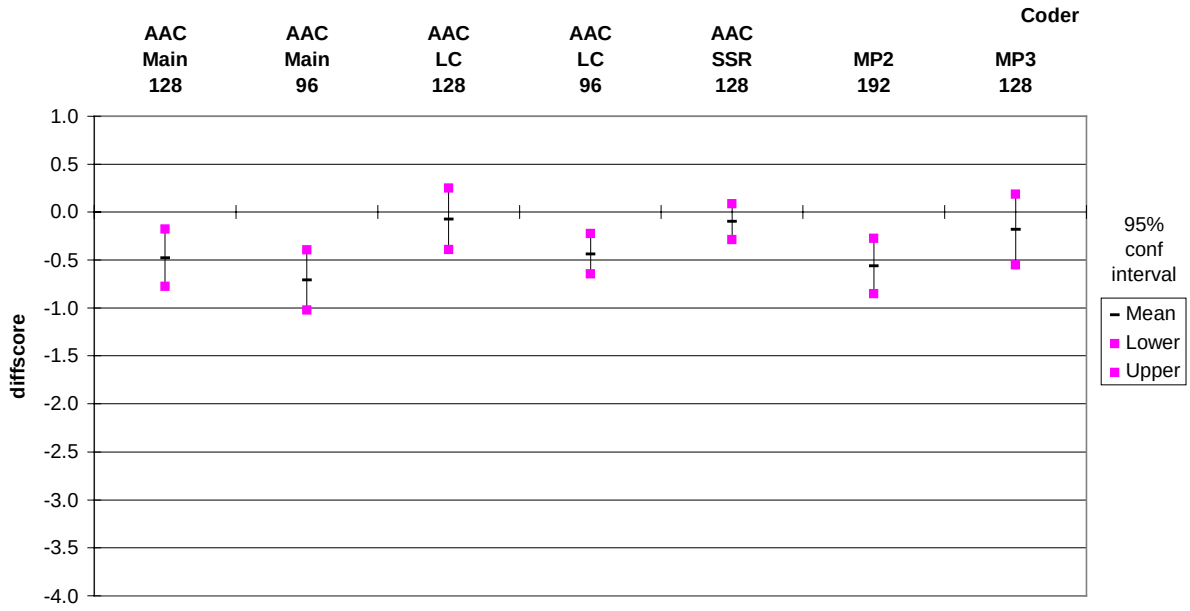
Pitch Pipe



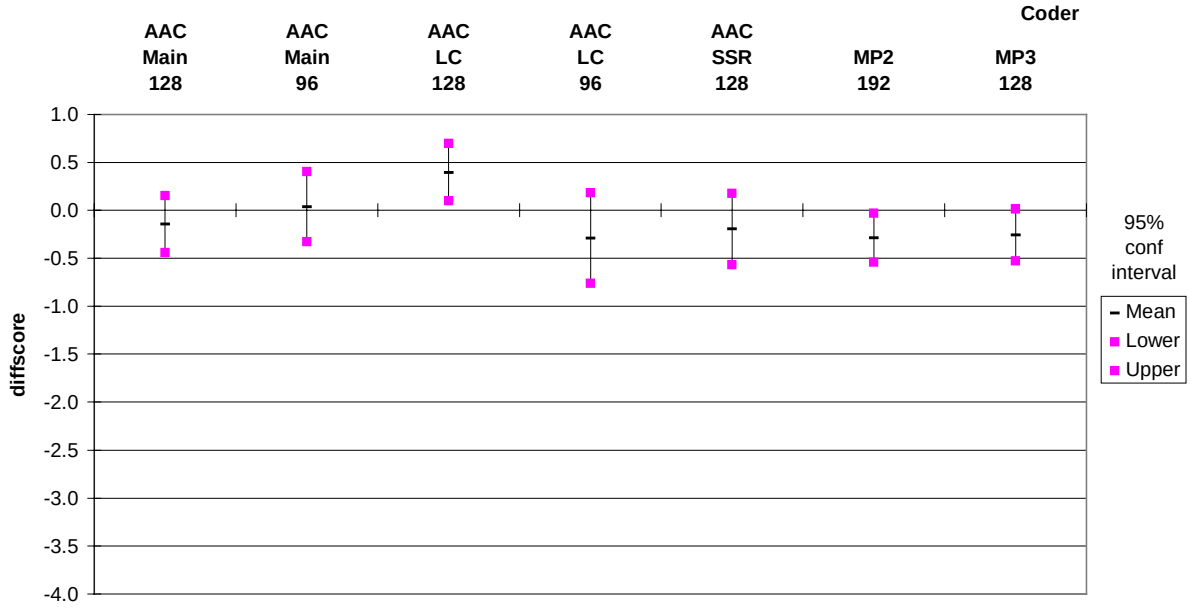
Glockenspiel



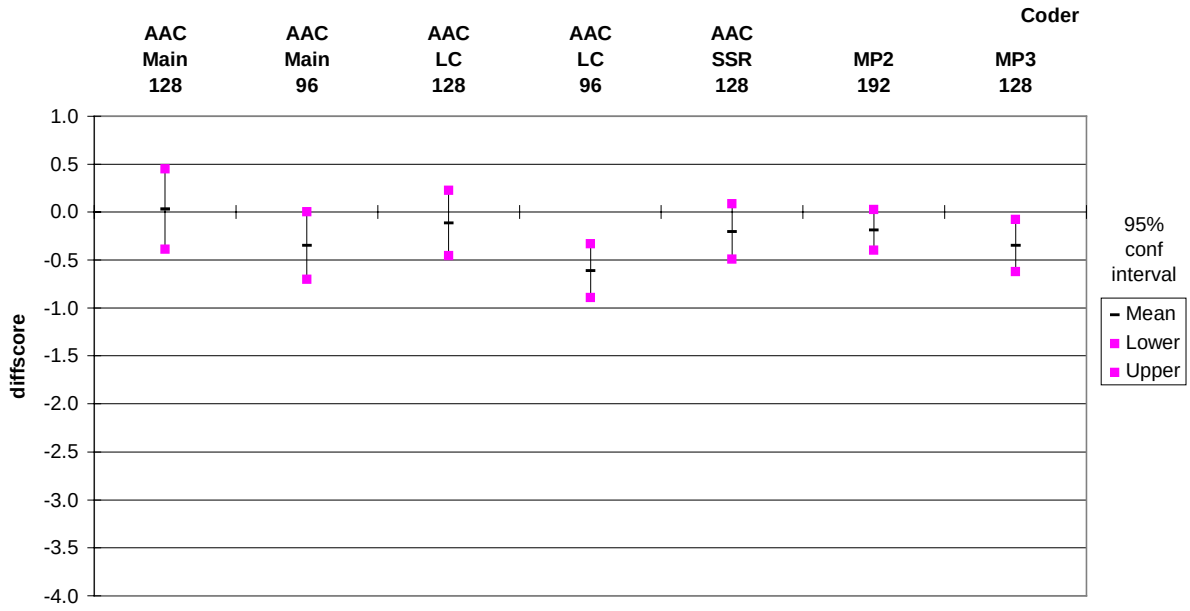
German Male Speech



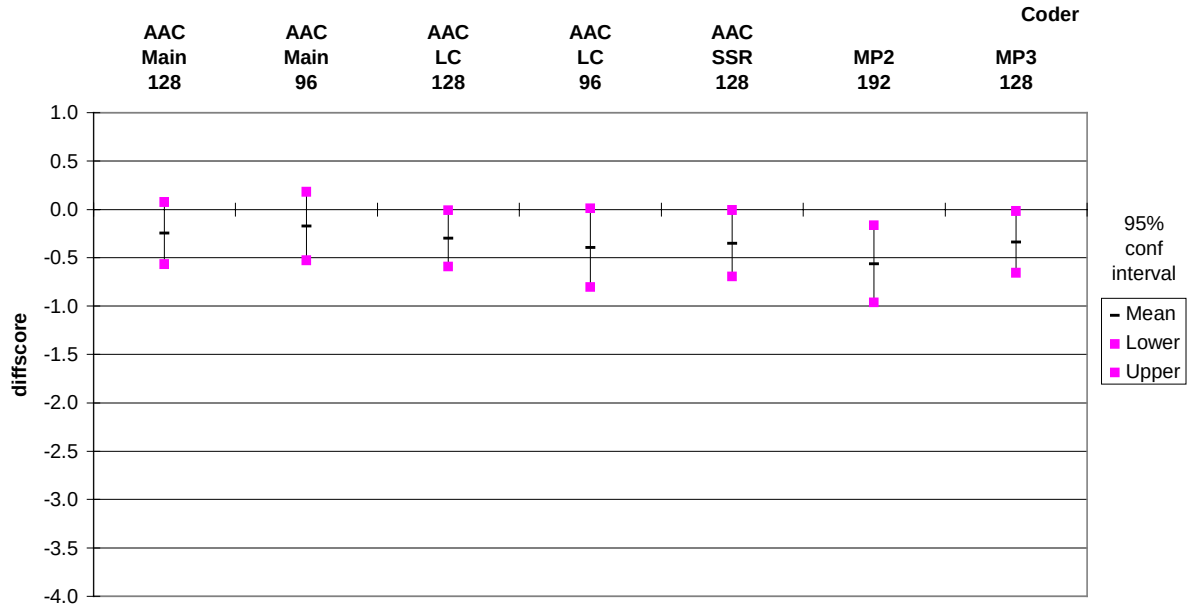
Suzanne Vega



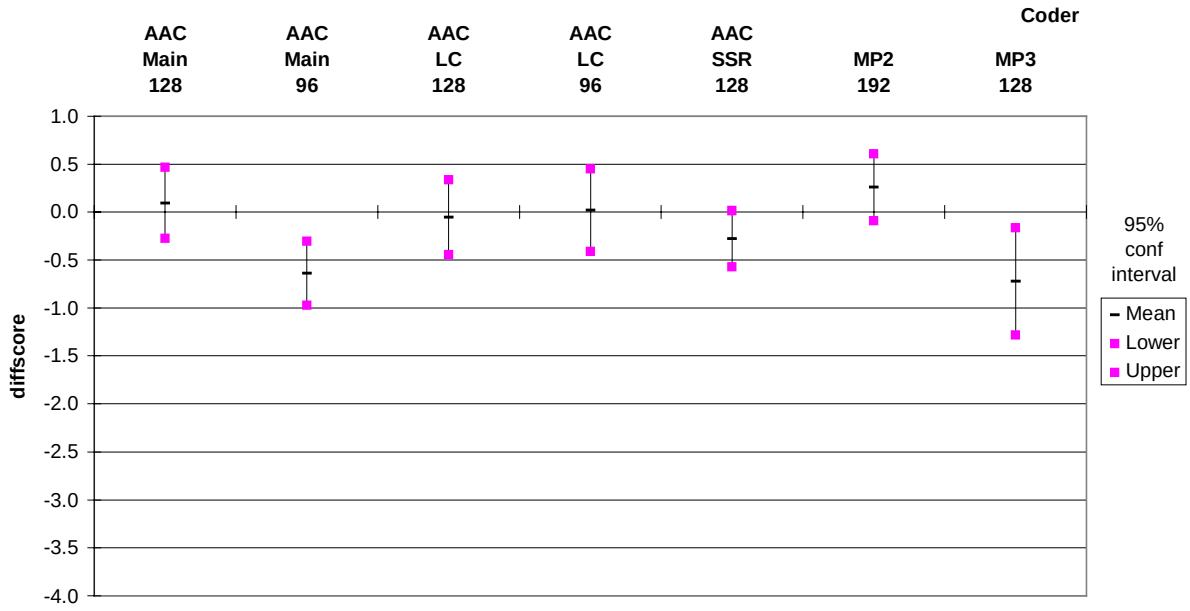
Tracy Chapman



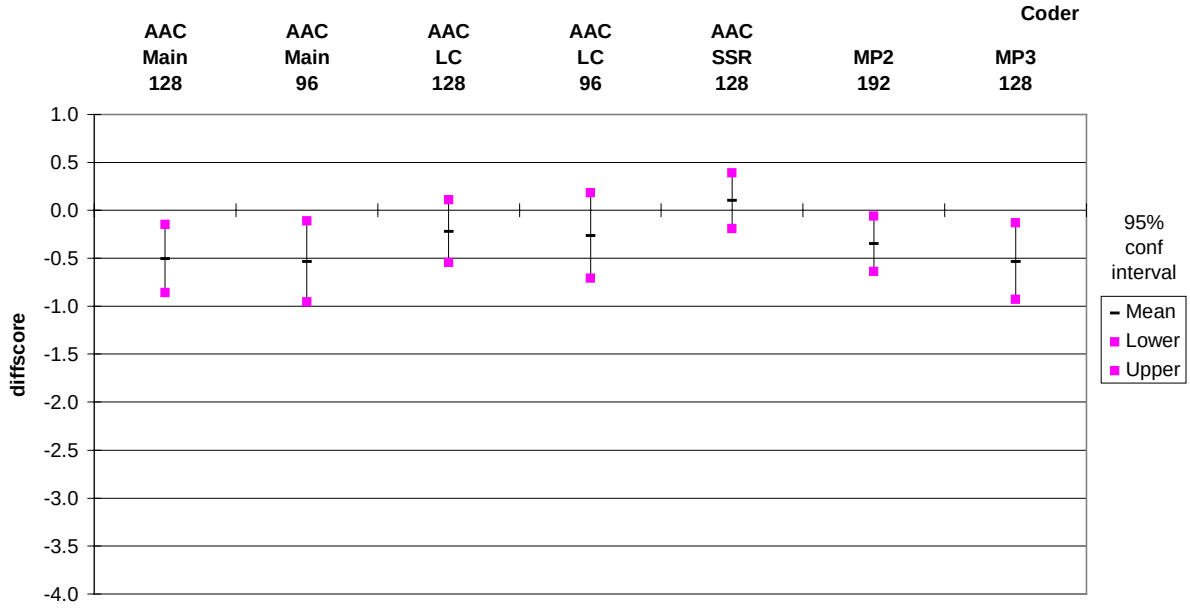
Ornette Coleman



Accordion & Triangle



Dire Straits



Annex 7. Test data for EBU “indistinguishable quality” criterion

Means and confidence intervals of the reference and test data, for each item by coder. This data is used to analyse the EBU criterion of “indistinguishable quality”.

Coder	Item	Condition	Lower	Upper
AAC Main 128	0	REF	4.6796	5.0658
		TEST	4.3086	4.7278
	1	REF	4.3806	4.9103
		TEST	4.3672	4.8600
	2	REF	4.5626	4.9465
		TEST	4.6182	4.9909
	3	REF	4.5718	4.9555
		TEST	4.5428	4.9936
	4	REF	4.9300	5.0245
		TEST	4.2127	4.7873
	5	REF	4.6723	5.0550
		TEST	4.5288	4.9076
	6	REF	4.4110	4.8800
		TEST	4.4074	4.9471
	7	REF	4.7128	4.9690
		TEST	4.3501	4.8408
	8	REF	4.4174	4.9280
		TEST	4.5655	4.9708
	9	REF	4.8374	5.0172
		TEST	4.1051	4.7404
AAC Main 96	0	REF	4.8847	5.0153
		TEST	3.4842	4.4249
	1	REF	4.6295	5.0705
		TEST	3.0077	3.9559
	2	REF	4.4959	5.0859
		TEST	3.8267	4.5643
	3	REF	4.6020	4.9616
		TEST	4.4554	4.8900
	4	REF	4.9387	5.0158
		TEST	3.9686	4.5678
	5	REF	4.4874	4.9398
		TEST	4.5195	4.9805
	6	REF	4.7614	4.9840
		TEST	4.2306	4.8149
	7	REF	4.6481	4.9974
		TEST	4.3873	4.9127
	8	REF	4.8832	5.0077
		TEST	4.0059	4.6123
	9	REF	4.7367	4.9997
		TEST	3.9842	4.6886
AAC LC 128	0	REF	4.6883	4.9663
		TEST	4.4492	4.8780
	1	REF	4.6941	5.0332
		TEST	3.7411	4.4953
	2	REF	4.9720	5.0098
		TEST	3.4409	4.2591
	3	REF	4.4473	4.9254
		TEST	4.5951	4.9958
	4	REF	4.5853	4.9693
		TEST	4.5084	4.9007
	5	REF	4.2665	4.7880
		TEST	4.8407	5.0047
	6	REF	4.4488	4.9876
		TEST	4.4187	4.7904
	7	REF	4.7832	4.9986
		TEST	4.3575	4.8243

Coder	Item	Condition	Lower	Upper
	8	REF	4.5540	4.9187
		TEST	4.3910	4.9726
	9	REF	4.6984	4.9380
		TEST	4.3492	4.8508
AAC LC 96	0	REF	4.7109	5.0891
		TEST	3.1356	4.2462
	1	REF	4.6895	5.0742
		TEST	3.0895	4.0651
	2	REF	4.9853	5.0420
		TEST	2.5833	3.6985
	3	REF	4.8700	5.0028
		TEST	4.2051	4.7131
	4	REF	4.9656	5.0071
		TEST	4.3458	4.7542
	5	REF	4.5843	4.9975
		TEST	4.1316	4.8684
	6	REF	4.8455	5.0364
		TEST	4.0974	4.5571
	7	REF	4.6920	4.9989
		TEST	4.1220	4.7780
	8	REF	4.3858	4.9142
		TEST	4.4056	4.9308
	9	REF	4.5843	4.9975
		TEST	4.1862	4.8683
AAC SSR 128	0	REF	4.4278	4.8541
		TEST	4.5415	5.0131
	1	REF	4.9313	5.0142
		TEST	3.2518	4.2300
	2	REF	4.8600	5.0491
		TEST	3.4671	4.2511
	3	REF	4.5237	5.0944
		TEST	4.4131	4.8233
	4	REF	4.8343	4.9839
		TEST	4.6599	4.9583
	5	REF	4.5830	4.9988
		TEST	4.3505	4.8404
	6	REF	4.7140	5.0132
		TEST	4.4549	4.8633
	7	REF	4.7393	4.9880
		TEST	4.2392	4.7881
	8	REF	4.7484	5.0425
		TEST	4.4003	4.8361
	9	REF	4.5161	4.9021
		TEST	4.6531	4.9650
MP2 192	0	REF	4.4282	4.9627
		TEST	4.1336	4.8845
	1	REF	4.7786	5.0396
		TEST	2.8133	3.8776
	2	REF	‡	‡
		TEST	1.9423	2.6941
	3	REF	4.8499	5.0410
		TEST	3.6737	4.4354
	4	REF	4.9300	5.0245
		TEST	4.1395	4.6877
	5	REF	4.8878	5.0031
		TEST	4.4274	4.8908
	6	REF	4.7832	4.9986

‡ In each of these cases, the listeners all correctly identified the Reference signal and assigned it the grade '5'. As a consequence, the standard deviation is 0 and the confidence interval does not exist.

Coder	Item	Condition	Lower	Upper
		TEST	4.5627	4.8464
	7	REF	4.8337	5.0299
		TEST	4.0048	4.7316
	8	REF	4.3106	4.8803
		TEST	4.7257	4.9834
	9	REF	4.8024	5.0340
		TEST	4.3338	4.8025
MP3 128	0	REF	4.9860	5.0049
		TEST	2.8165	3.9381
	1	REF	4.3127	5.1419
		TEST	2.1367	3.2724
	2	REF	‡	‡
		TEST	1.3105	1.9531
	3	REF	4.7118	4.9427
		TEST	4.2903	4.8643
	4	REF	4.6073	5.0199
		TEST	4.3783	4.8854
	5	REF	4.8096	5.0086
		TEST	4.4244	4.8756
	6	REF	4.8186	5.0269
		TEST	4.3495	4.7960
	7	REF	4.7332	5.0032
		TEST	4.2900	4.7737
	8	REF	4.5820	5.0817
		TEST	3.6775	4.5407
	9	REF	4.9133	5.0140
		TEST	4.0469	4.8167
codec_x	0	REF	4.4042	4.9049
		TEST	3.1954	4.3682
	1	REF	‡	‡
		TEST	1.6566	2.6252
	2	REF	‡	‡
		TEST	1.5060	2.3850
	3	REF	4.7628	5.0008
		TEST	3.7616	4.6929
	4	REF	4.8257	5.0106
		TEST	3.9156	4.6571
	5	REF	4.6831	5.0169
		TEST	4.2308	4.7692
	6	REF	4.8294	4.9888
		TEST	4.2095	4.7632
	7	REF	4.7178	5.0094
		TEST	4.0241	4.7123
	8	REF	4.4789	4.9484
		TEST	4.1107	4.8348
	9	REF	4.8438	5.0198
		TEST	3.9967	4.7487