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

ISO/IEC JTC1/SC29/WG11
MPEG98/N2276
July 1998

Title: **Report on the MPEG-4 audio NADIB verification tests**
Authors: **Catherine Colomes (CCETT), Caroline Jacobson (TERACOM), Eric Scheirer (MIT), Audio Subgroup**
Status: **Approved**

Summary

The MPEG-4 Audio coding tools covering 6kbit/s to 24kbit/s have undergone verification testing for an AM digital audio broadcasting application in collaboration with the NADIB(Narrow Band Digital Broadcasting) consortium. With the intent of identifying a suitable digital audio broadcast format to provide improvements over the existing AM modulation services, several codec configurations involving the MPEG-4 CELP, TwinVQ, and AAC tools have been compared to a reference AM system. It was found that higher quality can be achieved in the same bandwidth with digital techniques and that scalable coder configurations offered performance superior to a simulcast alternative.

1 Table of Contents

2 Introduction

The MPEG-4 Audio coding tools cover a bitrate range from 2 kbit/s to 64 kbit/s with a corresponding subjective audio quality that needs to be evaluated. It was recognised that the verification tests should first address applications that are potentially of great interest for users. Three main applications for this first round of MPEG-4 audio verification tests have been identified [1]:

- Internet Audio applications
- digital audio broadcasting on AM modulated bands (16 to 24 kbit/s) and
- speech applications

The NADIB (Narrow Band Digital Broadcasting) consortium proposed to carry out the MPEG-4 verification tests for digital audio broadcasting based on their proposals [2] [3] Two different sites offered to run the listening tests : TERACOM (Stockholm - Sweden) and CCETT (Rennes - France). The final results analysis was performed by MIT (USA).
The whole process was completed by the beginning of June, and it is the purpose of this document to describe the procedures that have been followed and to present the outcome of the tests.
All the specifications listed below refer only to the NADIB tests, the two other verification tests being handled in separate documents.

3 Context and test motivation

NADIB is a project under Eureka (EU 1559) focusing on digital audio broadcasting on AM modulated bands, like HF, MF and LF. The intention of this group is to define a world-wide usable standard for improving the existing analogue service in the above mentioned AM modulated bands. For this challenge the consortium wants to use latest technology in modulation schemes, channel coding and source coding. The CODEC-group, responsible for source coding aspects, had the opinion that the NADIB system would greatly benefit in applying the newest audio coding systems under development in the standardisation process of MPEG-4 Audio. Therefore the CODEC group stated a proposal for defining an MPEG-4 Audio profile for Digital Broadcasting in the AM band that has been integrated in the MPEG-4 Applications Document [4]. NADIB offered to conduct verification tests of the MPEG-4 audio tools in line with the NADIB requirements.

The motivation for having these tests was to get an impression of the coding efficiency and the coding gain of the new MPEG-4 system, compared under different test conditions, especially in the scaleable vs. not scaleable mode. A scaleable system is of high interest since it allows a broadcasting system to be designed in a way that it offers full quality under good reception conditions (full bitrate available, i.e. core + enhancement layer) while still having meaningful output under bad error conditions (only core coder data available). Being a realistic condition for the narrowband digital audio broadcasting, this listening test was to be done with a bitrate of 6 kbps for the core coder and a total bitrate of 24 kbit/sec for the bitstreams (in monophonic mode). This means that the bitrate of the enhancement layer (or a separate coder working simultaneously) is 18 kbit/sec.

The comparisons of interest were defined as follows [5]:

1. to compare the coding efficiency of different possible core coders: MPEG-Narrowband CELP as a speech coder and Twin-VQ as a generic audio coder. Furthermore, G723.1 is added as an anchor point.

2. to evaluate the advantages/disadvantages of a system with two layers versus a system using only one layer. Therefore the unscaled AAC codec is compared to the scaleable versions at the same total bitrate (24 kbps)
3. to compare scaleability against simulcast. In the simulcast mode, as many bit streams as potential decoders are broadcast in parallel. This solution makes the decoding process less complex since only one decoder at a time has to be used in order to get the desired quality. On the other hand, upper layers can not take advantage from lower layers and the coding is expected to be less effective. Therefore it is desirable to compare, at a given bit rate, the audio quality for both solutions

The deadline for the encoding process was April 1st, 1998. The test was conducted at CCETT (France) and Teracom (Sweden) in May 1998. For details on the test schedule see ANNEX 1.

4 Codecs Under Test

4.1 Test Overview

During the Tokyo meeting, where the test activities were finalised, the following decisions were taken [5]:

- the test has to be divided in two groups:
- *Test A* contains narrowband CELP (NB-CELP), TwinVQ and G723.1. The reference signals for this group are the 8 kHz originals
- *Test B* contains the wideband CELP (WB-CELP), the higher rate audio coders and perfect AM. The reference signals for this test are bandlimited originals with a sampling rate of 24 kHz. The bandwidth of the reference will be the same as the bandwidth of the coder offering the highest bandwidth
- All coders operate in mono mode. The stereo mode is not included in this MPEG-4 test, but it could be scheduled later on.
- All coders operate in fixed bitrate mode. A maximum short time buffer of 6144 bytes is allowed (this corresponds to the max. bit reservoir of AAC)
- All coders are tested with speech and music items, because both of them are relevant for audio broadcasting

It should be noted that in MPEG standards only the decoder is normative and that the MPEG-4 codecs supplied for these tests are developmental and further optimisation is expected. It must be stressed that some of the coders in the test are speech coders which were not designed for music which is present in several items used in this test.

The codecs which were tested are listed below :

Test & #codec	Codec	Delivered by	Sampling rate of codec	Total bitrate (layer bitrate) in kbit/s	Estimated bandwidth(4)	Sampling rate of reference
A1	Narrowband-CELP	NEC	8 kHz	6	3.5	8 kHz
A2	Twin-VQ	NTT	24 kHz	6	3.5	8 kHz
A3	G.723.1 (1)	CCETT	8 kHz	6.3	3.5	8 kHz
B1	Wideband -CELP (2)	Philips	16 kHz	18.2	7.5	24 kHz
B2	AAC pure (2)	FhG	16 kHz	18	6.5	24 kHz
B3	AAC pure (3)	FhG	24 kHz	24	7.5	24 kHz
B4	AAC scal. w. CELP core	FhG/NEC	24 kHz	24 (6+18)	7	24 kHz
B5	AAC scal. w. TwinVQ core	FhG/NTT	24 kHz	24 (6+18)	6	24 kHz
B6	MPEG-2 Layer III (1)	FhG	24 kHz	24	6	24 kHz
B7	perfect AM (1)	Deutsche Welle	N/A	N/A-	-3db at 2.4 -50dB at 5.3	24 kHz

(1) reference anchor

(2) for scaleability Vs simulcast comparison

(3) for two layer Vs one layer comparison

(4) estimated with CoolEdit

Parallel to the test the bitrate and conformance of all coders was verified. For details of the analysis see ANNEX 4.

4.2 Codec Details

4.2.1 Narrowband CELP

The narrowband CELP coder used in this test was improved compared to the VM Software available at that time. The improvements were only done in informative parts. In the Decoder, an improved postfilter was used. The Encoder was improved as described in M3357 and M3502 (contributions to the Tokyo meeting 1998). The coder used mode VIII (vector quantizer with multipulse excitation) as described in FCD available at that time) and a bitrate of 6.0 kbps. The frame length was 20 msec and the delay was 25 msec.

4.2.2 TwinVQ

The TwinVQ under test is the coder newly designed as a result of the AAC-TwinVQ convergence work, whose specifications are described in the FCD. It quantizes a part of 1024/128 point MDCT coefficients for 24 kHz sampling rate input, and can be directly plugged into the AAC scaleable system. It has no increase of delay nor complexity due to the scaleable operation, since there is no need to execute a up/down sampling process in the encoder or the decoder.

4.2.3 G723.1

The G723.1 is a speech encoder recommended by ITU-T for multimedia communication at 5.3 and 6.3 kbps. In this test the 6.3 kbps version was used. This encoder was optimized for encoding speech signals with a high quality for a limited amount of complexity. The frame length is 30 msec with an additional look ahead of 7.5 msec, resulting in a total algorithmic delay of 37.5 msec.

4.2.4 Wideband CELP

The Wideband CELP operated in mode III (scalar quantizer and regular pulse excitation, 16 kHz sampling rate). A fixed bitrate of 18200 bps was used. Encoded bandwidth was 7.5 kHz, delay was 18.75 ms

4.2.5 AAC pure

The AAC coders used in this test were MPEG-2 AAC Low Complexity profile encoders according to ISO/IEC 13818-7

4.2.6 AAC scaleable with CELP core

This coder is a 2 layer scaleable coder. The lower layer is the MPEG-4 Narrowband CELP core coder as available in the MPEG-4 VM at the time of the bitstream delivery deadline. The higher layer is a MPEG-4 AAC layer with the PNS tool enabled.

4.2.7 AAC scaleable with Twin VQ core

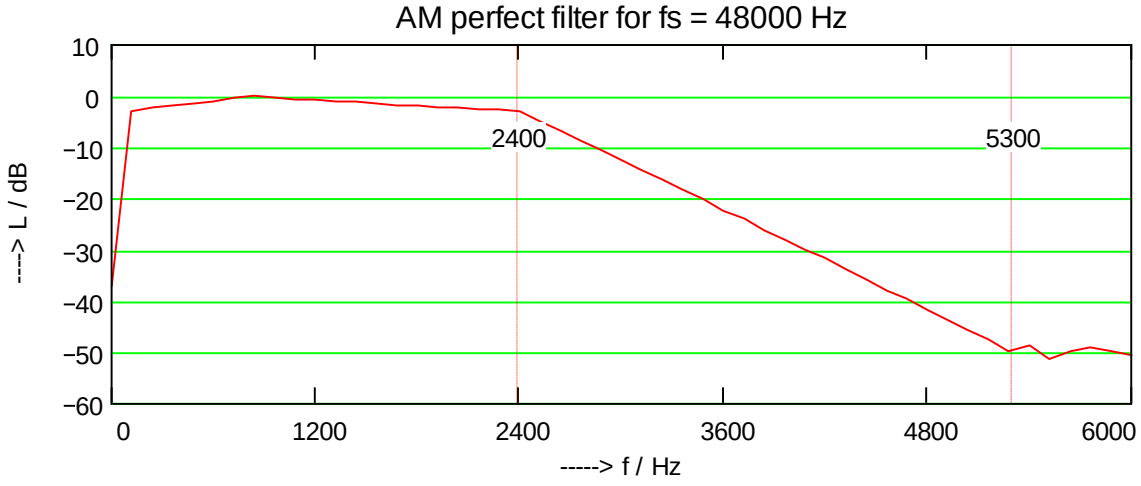
This coder is a 2 layer scaleable coder. The lower layer is the MPEG-4 TwinVQ as described above. . The higher layer is a MPEG-4 AAC layer with the PNS tool disabled.

4.2.8 MPEG-2 Layer 3

The Layer 3 coder operates at 24 kHz sampling rate according to ISO/IEC 13818-3.

4.2.9 Perfect AM

Perfect AM is simulated using a bandpass filter with the following characteristics: -50dB at 24Hz, -3dB at 73Hz, -3dB at 2400Hz and -50dB at 5300Hz.



This characteristic is based on a measurement of consumer AM receivers by Deutsche Welle showing that the -3dB point of the lowpass filter is at or (mainly) below 2.4 kHz for most

receivers. At the same time, most receivers reach -50dB at 5.3 kHz. It should be noted that no other distortions, which might be caused by the AM modulation scheme, were taken into account in this simulation of AM.

5 Test Material

A call for new test material (specifically speech items in different languages) was sent out during winter 97. This resulted in a total of more than 140 items all together - speech and music. Out of these, 51 representative items have been selected to be worth considering. (see the list in **ANNEX 5**). This selection was made by Martin Dietz of FhG and Mr. Schall of Deutsche Welle. Finally the files were the sent to Samsung for cut and editing process before being made available to the selection panel for the final selection.

5.1 Selection Panel

The process of identifying and selecting the most critical programme items to be used in the formal tests was delegated to a selection panel and carried out at Swedish Radio. The selection panel was comprised of :

- JY Leseure (CCETT - France)
- L Mossberg (Swedish Radio)
- W Schäfer (Sony)
- N Schall (Deutsche Welle)

For the final selection of the mono test excerpts it was proposed to have half of the selected excerpts as speech excerpts and the other half as music excerpts. For speech excerpts there should be at least English speech (male & female) and the native language of the test site (French/Swedish, male & female).The tasks of the selection panel and the characteristics of the excerpts were defined as listed below :

- 12 mono test excerpts should have been selected for speech and complex signals, distributed as follows:
    - 8 common items, to be used in both test sites
    - 2 French items, to be used at CCETT only
    - 2 Swedish items, to be used at Teracom only.
- Moreover, the 8 common items should have to include:
    - 3 music items
    - 3 music/noise&speech
    - 2 English speech
- Excerpts should have to be critical for **each** of the codecs under test. In other words the excerpts, when encoded, should have presented clearly audible impairments with different characteristics
- The selection should have been done without having information about the identity of the codecs
- Items chosen as critical at low bitrates should have had to be included also in the high bitrate and vice versa, i.e. the same excerpts should have had to be used in both tests.
- Training items, different from those used in the tests, should have had to be selected. Training items could have been different for each test. For test A, 2 music + 2 speech, should have been selected, while for group B more than 4 items could have been useful (maximum 8) to demonstrate the artefacts that subjects could expect to listen during the test.

- if the subjective loudness would have been found to be different from that of the reference or coded versions, a level for the 'perfect AM processed version' should have been adjusted
- advices should have been given in advance to Swedish Radio about the technical equipment to be used in the selection process

The selection panel was invited to follow the instructions above, unless there was any evidence for a need for changing the distribution of kind of items.

5.2 Chosen Items

After a week of work, the selection panel recommended a set of 10 items for the A and B tests, and also suggested 4 different orders of playing sequences. The selection panel also recommended specific items to be used during the training phase of the listeners. Additional details on this selection process can be found in **ANNEX 6.**

5.2.1 Items for the A and B tests

The numbers listed below refer to the items numbers of **ANNEX 5**
The following program items were used in the test. Item36 and item50 were specific for Teracom while CCETT used item13 and item28 instead. The other eight items were used at both test-sites.

Item	Content
Item2	Male voice (English)
Item10	Female voice (English)
Item13	Female voice (French)
Item20	Speech + music
Item21	Pop music
Item22	Folk music
Item28	Male voice (French) + music
Item35	Music + noise
Item36	Female voice (singing)
Item38	Classic music
Item44	Speech + noise
Item50	Speech + noise (Swedish)

5.2.2 Training Items

The training items, proposed by the selection panel, were chosen to be different from the items used during the formal tests. The associated codecs were proposed by both TERACOM and CCETT in order to cover the whole range of quality that will be encountered during the tests.

For the training the same items were used at both test sites. These items are as follows:

Training for the A test :

Item	Contents	Codec
Item3	Male voice (English)	**A3**
Item12	Female voice (English)	**A1**
Item39	Speech + music (Japanese)	**A2**
Item48	Speech + noise (German)	**A1**

Training for the B test:

Item	Contents	Codec
Item3	Male voice (English)	**B7**
Item12	Female voice (English)	**B6**
Item23	Classic music	**B4**
Item31	Speech, male + female (French)	**B1**
Item33	Music + noise	**B5**
Item39	Speech + music (Japanese)	**B3**
Item48	Speech + noise (German)	**B2**

6 Test Methodology

The NADIB consortium proposed to have mainly non-expert listeners involved in the tests. The grading in degradation levels might have been inappropriate for non experienced listeners. Thus, the assessment method defined in the ITU-R Recommendation BS 1284 (former BS.562.3) was used [6]. BS 1284 uses the following 5-grade scale:

	BS 1284 (BS.562.3) Quality scale
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

The quality scale was used as a continuous scale with a resolution of one decimal.

7 Test Stimuli

Each condition (codec, item) was assessed using a double stimulus presentation. In the double stimulus test with a grading on a quality scale, two audio stimuli, signal "Ref" and signal "A" are presented to the listeners. Only signal "A" is assessed by the listeners, signal "Ref" serves as an indication of the optimum quality for the considered item. In consequence, the listeners are invited to score the signal "A" comparatively to signal "Ref" using a dedicated quality scale. The stimuli were presented to the listeners according to the sequence shown in figure 1. They were pre-recorded on a digital storage media (DAT tape or Optical Disk) together with aural announcements to help the listeners to keep the track.

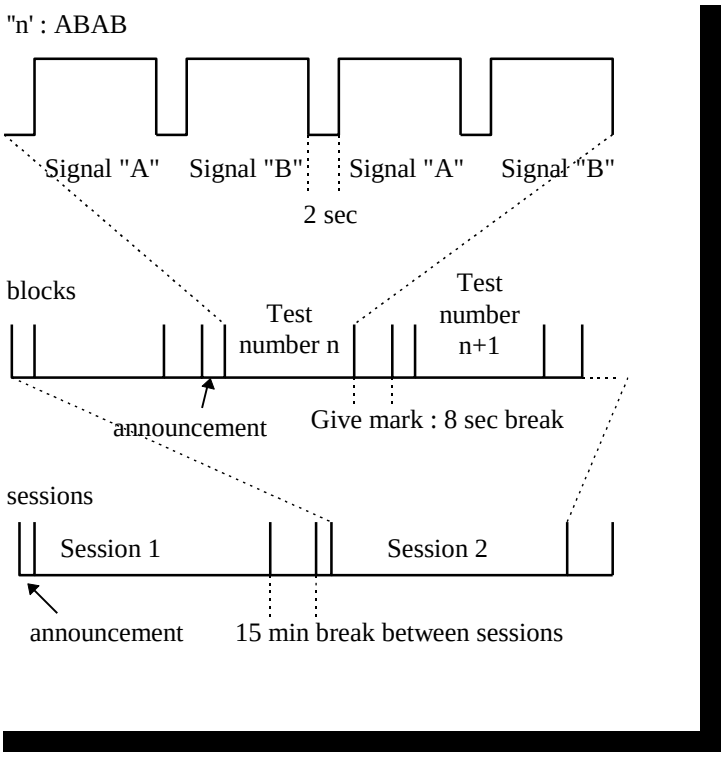


Fig. 1 : Protocol of the double stimuli test graded on a quality scale.

8 Test sessions

To ensure fatigue did not affect the results, A and B tests were split into sessions of approximately 20-25 minutes. That is two sessions for the A test and 5 sessions for the B test, resulting in 7 sessions per listener.

Starting from the Selection Panel recommendations, a "pseudo-randomisation" of the test stimuli was applied to minimise the number of times each codec configuration occurred in a test session, and therefore to mix the audio quality throughout the test. As far as it was possible, the "pseudo-randomisation" was chosen in order to avoid too many repetitions of a test item in the same session. 3 repeated trials were included in each "pseudo-randomisation" of the A test, and 7 in each "pseudo-randomisation" of the B test in order to check the reliability of the listeners.

Finally, both test sites agreed on 4 different "pseudo-randomisations" per test A and B (see **ANNEX 8**). For more details on the test organizations, see ANNEX 2 and ANNEX 3.

9 Data Analysis

The aim of this test was to answer the following questions with respect to seven MPEG-4 codecs in two tests:

1. Are the listeners reliable; i.e., are their responses consistent?
2. How do the results from the two test-sites (Teracom SE and CCETT FR) compare?
3. How is the performance of the MPEG-4 codecs with respect to the anchor conditions?
4. How does the performance of the codecs vary with programme item?
5. What is the relative ranking of the codecs tested?

6. How is the performance of 1-layer AAC coding compared to scaleable (2-layer) coding at the same total bitrate?
7. How is the performance of scaleable coding compared to multicast (base layer only) coding?
8. What is the performance of the scaleable codecs compared to perfect analogue AM transmission?
9. What is the performance of AAC coding compared to MPEG-4 WB CELP, both at 18 kbps?

Two tests were conducted. In the first, a test of low-bitrate codecs operating on 8 kHz signals compared MPEG-4 Twin-VQ coding at 6 kbps to MPEG-4 Narrowband CELP (NB-CELP) at 6 kbps. The ITU-T standard G.723.1 ADPCM codec at 6.3 kbps was used as a reference.

In the second, 5 codecs operating at medium bitrate on 24 kHz signals were compared. The coders were: MPEG-AAC at 24 kbps and at 18 kbps (AAC-24 and AAC-18), MPEG-4 Wideband CELP (WB-CELP) at 18.2 kbps, scaleable AAC with a CELP code at a total of 24 kbps (AAC/CELP), and scaleable AAC with a Twin-VQ code at a total of 24 kbps (AAC/TwinVQ). MPEG-2 Layer III coding at 24 kbps (MP3), and a simulation of perfect (noiseless) AM radio transmission were used as references.

This paragraph details the statistical analysis of the listening-test data. Tables referenced in the following text have been gathered in Annex 9.

9.1 Data Receipt and Verification

Data collected at the two test sites were received by MIT on 23 May 1998. The data were provided in the form of several Microsoft Excel spreadsheets. The 'overall' spreadsheet data were re-written to disk as tab-delimited text files with the same format as the spreadsheets. A set of PERL programs was then written to unroll the textual spreadsheets into a common format, where each line contained one rating for one listener in one trial. This format contains numerous rows with the following organisation:

SITE SUBJECT EXPERT BLOCK TEST SESSION TRIAL ITEM CODEC SCORE

The SITE variable indicates at which site the rating was recorded; the SUBJECT variable indicates the subject number; the EXPERT variable indicates whether the subject was an expert listener, a non-expert, or when this status was not known; the BLOCK variable indicates which block of trials the test was conducted in; the TEST variable indicates whether the trial was a low-bandwidth (8 kHz) or high-bandwidth (24 kHz) trial; the SESSION variable indicates the testing session in which the trial took place; the TRIAL variable indicates the number of the trial within the block; the ITEM variable indicates the programme item under test; the CODEC variable indicates the codec under test; and the SCORE variable indicates the subject rating for the trial. The EXPERT, BLOCK, SESSION, and TRIAL variables were not used for analysis.

The data were checked for completeness. Except as noted, this and all other analysis was conducted using the SPSS Version 7.5 for Windows statistical package. The CCETT data, when read in, consisted of 2497 cases, with each subject participating in 11 trials per codec, except for subject C7, who did not test codecs A1, A2, and A3. Subject C7 was therefore removed from the data set, leaving 2420 cases in the CCETT data set. In this set, each subject participated in 11 trials per codec; each item for each codec was tested 22 times, except for the 10 item/codec pairs for which there were repeated data, for which there were 44 tests.

The Teracom data consisted of 3960 cases, with each subject participating in 11 items per codec. Each item for each codec was tested 36 times, except for the 10 item/codec pairs for which there were repeat data, for which there were 72 trials.

9.2 Listener Reliability

Another PERL script was used to search through the textual spreadsheet data and locate the repeated codec/item pair. A data file was created with one case for each such repeated test, in the following format:

SITE SUBJ ITEM CODEC SCORE1 SCORE2

The SITE, SUBJ, ITEM, and CODEC variables are as above; the SCORE1 and SCORE2 variables give the score of each of the two repeated trials for this item and codec.

This data file had 580 cases when read in, 10 repeated trials for each of 58 subjects.

The following heuristic was developed to evaluate subjects for reliability. For each subject, the mean and 95% confidence interval of the score difference (SCORE1-SCORE2) would be calculated. If, for any subject, the confidence interval extended beyond [-1,1], that subject would be eliminated. Thus, for subjects retained, one has a high confidence that their true scores on other trials are within 1 rating level of the judgements presented.

The means and confidence ratings for each subject on this criterion are shown in Table 1 (Annex 9). Four subjects were eliminated as unreliable using this measure: C23, T1, T14, and T25. One should note that although the [-1,1] confidence interval is an arbitrary cut-off point, any cut-off level between 0.86 and 1.05 would have eliminated the same four subjects, indicating that these subjects actually had quite different behaviour from the others.

Eliminating these subjects leaves us with 5940 cases: 54 subjects with 110 trials each. From this, one of the repeated pair of trials for each repeated test was eliminated at random, to ensure that the ANOVAs below were balanced. This final elimination provides the analysis data set of 5400 cases: 54 subjects, with 10 codecs and 10 items per codec each.

9.3 Test site comparison

A factorial ANOVA was computed to compare the results of the testing at each of the two test sites. The result is shown below:

ANOVA[a,b]

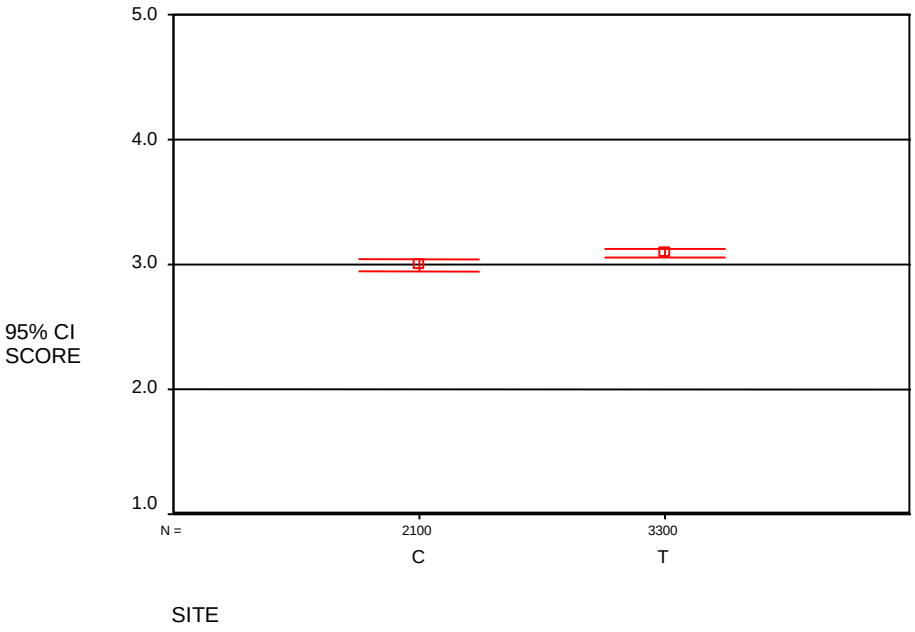
			Unique Method				
			Sum of Squares	df	Mean Square	F	Sig.
SCORE	Main Effects	(Combined)	1792.647	16	112.040	230.787	.000
		SITENUM	7.972	1	7.972	16.422	.000
		CODECNUM	1765.419	9	196.158	404.056	.000
		ITEMNUM	19.255	6	3.209	6.610	.000
	2-Way Interactions	(Combined)	917.947	69	13.304	27.403	.000
		SITENUM * CODECNUM	11.350	9	1.261	2.598	.005
		SITENUM * ITEMNUM	3.919	6	.653	1.345	.233
		CODECNUM * ITEMNUM	902.678	54	16.716	34.433	.000
	3-Way Interactions	SITENUM * CODECNUM * ITEMNUM	19.942	54	.369	.761	.901
	Model		2814.151	139	20.246	41.703	.000
	Residual		1767.118	3640	.485		
	Total		4581.270	3779	1.212		

a. SCORE by SITENUM, CODECNUM, ITEMNUM

b. All effects entered simultaneously

One can observe several points about this analysis. First, there is a significant effect of test site, which is further shown in Fig. 1. Scores were consistently higher at the Teracom test site than at the CCETT test site. Thus, the results from the two test sites must be calculated independently. Also, there was a significant interaction between test site and codec, so when analysing the results codec-by-codec, the two test sites must be calculated independently. However, there was no significant interaction between the test site and programme item, and no three-way interaction, so when analysing the results item-by-item, the two test sites may be pooled and considered together.

Fig.1: Comparison of results at the two test sites



9.4 Evaluation of codecs

9.4.1 Codec-by-codec

Means and confidence intervals for each of the codecs at each of the test sites were computed, to evaluate their overall performance. These results are shown in Table 2, and graphically in Figures 2 and 3 .

Additionally, the student tables for both sites and both test (i.e. A and B) have been computed, together with the ranking of codecs based on the mean grades. From these data it is possible to build the NSSD (Next Statistically Significant Difference) matrixes for each case :

Student Matrix		
	NB-CELP	G723.1
Twin VQ	0	0
NB-CELP		0.0093

Ranking	Twin VQ	NB-CELP	G723.1
mean grade	1.76	2.55	2.71
NSSD	NB-CELP	G723.1	-

Test A at Teracom

Student Matrix						
	Perfect AM	AAC-18	MPEG1-LIII	AAC/TWVQ	AAC/NBCELP	AAC-24
WB-CELP	0	0	0	0	0	0
Perfect AM		0	0	0	0	0
AAC-18			0.00083	0.000021	0	0
MPEG1-LIII				0.344	0.00021	0
AAC/TWVQ					0.0042	0
AAC/NBCELP						0

Ranking	WB-CELP	Perfect AM	AAC-18	MPEG1-LIII	AAC/TWVQ	AAC/NBCELP	AAC-24
mean grade	2.37	2.83	3.28	3.49	3.55	3.72	4.13
NSSD	Perfect AM	AAC-18	MPEG1-LIII	AAC/NBCELP	AAC/NBCELP	AAC-24	-

Test B at Teracom

Student Matrix		
	NB-CELP	G723.1
Twin VQ	0	0
NB-CELP		0.094

Ranking	Twin VQ	NB-CELP	G723.1
mean grade	2.01	2.66	2.86
NSSD	NB-CELP	-	-

Test A at CCETT

Student Matrix						
	Perfect AM	AAC-18	AAC/TWVQ	MPEG1-LIII	AAC/NBCELP	AAC-24
WB-CELP	0	0	0	0	0	0
Perfect AM		0.00102	0	0	0	0
AAC-18			0.00001	0	0	0
AAC/TWVQ				0.0661	0	0
MPEG1-LIII					0.2669	0
AAC/NBCELP						0

Ranking	WB-CELP	Perfect AM	AAC-18	AAC/TWVQ	MPEG1-LIII	AAC/NBCELP	AAC-24
mean grade	2.24	2.79	3.07	3.43	3.58	3.67	4.13
NSSD	Perfect AM	AAC-18	AAC/TWVQ	AAC/NBCELP	AAC-24	AAC-24	-

Test B at CCETT

The student matrixes show statistically significant difference between codecs for values below 0.05. For instance, for test A at Teracom, all codecs are different while for the same test at CCETT NB-CELP and G723.1 are statistically not different.

Averaged over all items used in the test, these results show the following:

- in the 8 kHz test, G.723.1 and NB-CELP were superior to Twin-VQ at both sites. They were statistically equivalent at CCETT but statistically different at Teracom.
- in the 24 kHz test at Teracom, AAC-24 was the best coder, and AAC/CELP was second in position.. MP3 was statistically equivalent to AAC/VQ but worse than AAC/CELP. Then, in order, AAC-18, Perfect AM, and WB-CELP, with significant differences at each step.
- in the 24 kHz test at CCETT, AAC-24 was the best coder. AAC/CELP and MPEG-2 LIII were statistically equivalent but worse than AAC-24. MPEG-2 LIII and AAC-/TWVQ were statistically equivalent. All four of these were better than AAC-18. All five of these were better than AM. WB-CELP was significantly worse than the other codecs.

9.4.2 Item-by-item

The data were pooled and examined item-by-item for each codec. These results are shown in Table 3, and graphically in Figures 4-15, and again for the transpose in Figures 16-25. One can analyse these data by tabulating the pairwise comparisons for each pair of codecs within each.

8 kHz test	Twin-VQ	G.723.1	NB-CELP
Twin-VQ		10	8
G.723.1	1		0
NB-CELP	2	1	

24 kHz test	AAC scal/ CELP	MPEG-2 Layer III	Perfect AM	AAC-24	WB-CELP	AAC scal/ Twin-VQ	AAC-18
AAC/CELP		0	0	7	0	0	0
MPEG-2 Layer III	0		0	8	0	0	0
Perfect AM	8	6		11	0	7	6
AAC-24	0	0	0		0	0	0
WB-CELP	9	7	6	11		8	8
AAC/Twin-VQ	1	0	0	9	0		0
AAC-18	8	6	2	12	2	5	

To read these charts, read down from a coder to see how on many items it was superior (out of 12 possible) to each other coder. For example, Twin-VQ was superior to G.723.1 on 1 item, and to NB-CELP on 2 items. Codecs with large values in their column performed well, and codecs with large values in their row performed poorly relative to the other codecs.

In general, for the 8 kHz test, there was little difference between NB-CELP and G.723.1 , and TwinVQ was significantly worse than these. For the 24 kHz test, AAC-24 performed significantly better than the other codecs. There were no programme items for which any codec was superior to AAC-24. After that, AAC/CELP, MP3, and AAC/VQ give very similar

performance. WB-CELP gave the worst performance of these codecs, and Perfect AM was next-to-worst.

9.5 Test results

The analytic data presented above give sufficient information to address the questions posed in the test plan. They are presented each in order.

9.5.1 Listener Reliability

«Are the listeners reliable; i.e., are their responses consistent?»

As discussed in Section IV, using the heuristic developed there, 54 of 58 listeners gave consistent responses and were included in the test after post-screening.

9.5.2 Test site comparison

«How do the results from the two test-sites (Teracom SE and CCETT FR) compare?»

As discussed in Section V, there was a main effect of test site, and an interaction between test site and codec. However, there was no interaction between test site and programme item, and no three-way interaction. Thus, the analysis in this Section was conducted separately by test site for the codec-by-codec evaluation, and pooled for the item-by-item evaluation.

9.5.3 Performance of codecs

«How is the performance of the MPEG-4 codecs with respect to the anchor conditions?»

The figures below summarise the relative performance of the MPEG-4 codecs obtained at the two test sites. Please note that the Layer 3 coder used was a MPEG-2 Layer 3 coder, although it is referenced as MPEG-1 Layer 3 in the figures.

Figure 2: Codec-by-codec result at CCETT

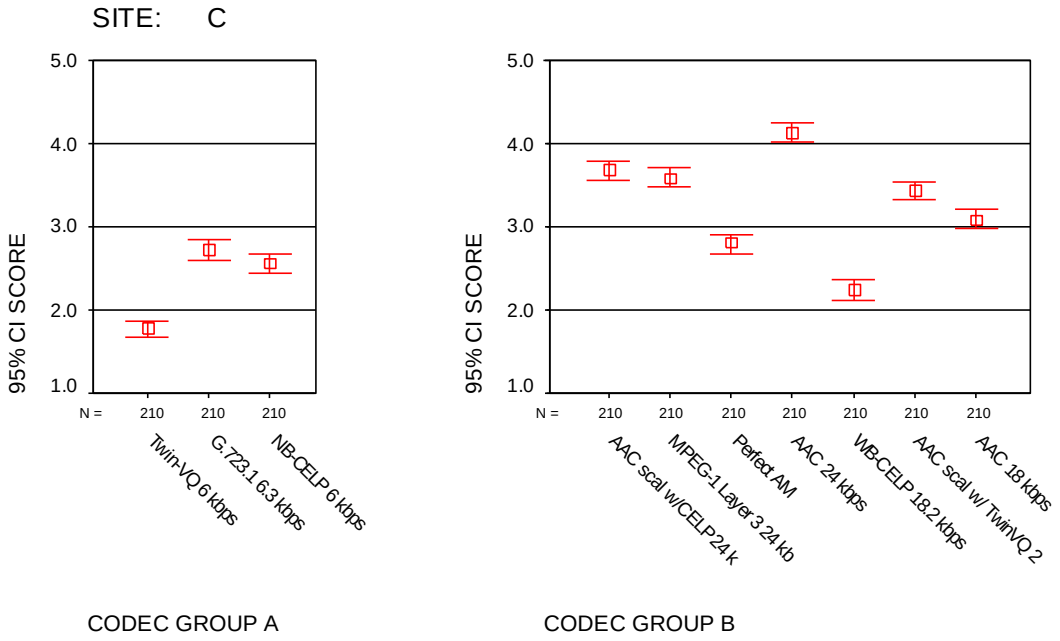
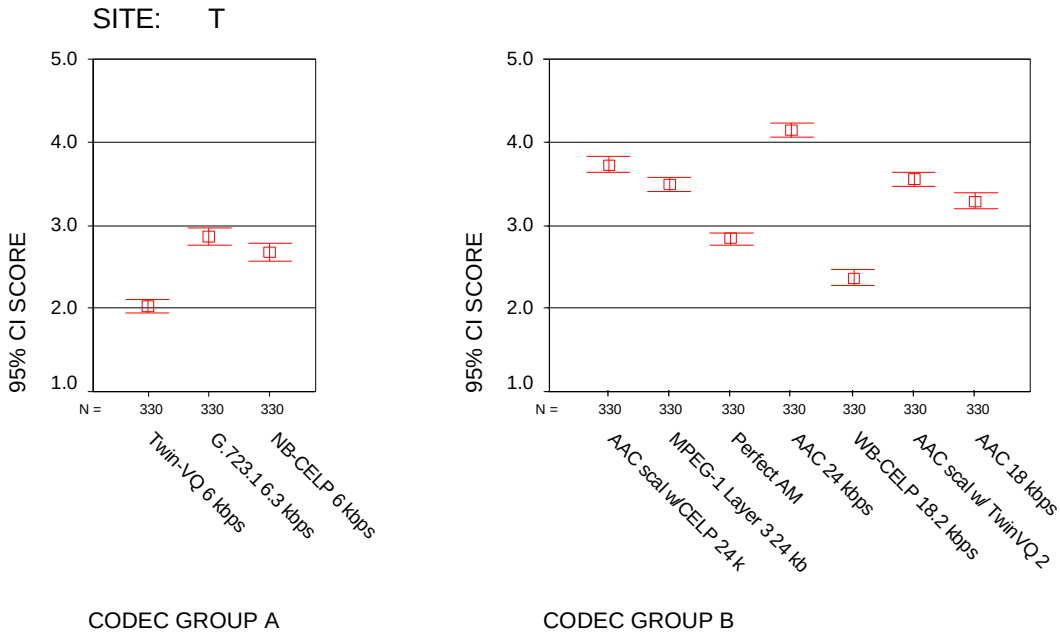


Figure 3: Codec-by-codec result at Teracom



In these diagrams, all different source items have been pooled to give an overall indication about the performance of each codec. However, as it is confirmed by the results of the ANOVA, the performance of the individual codecs is dependent on the source material being used, such as clean speech, speech with background noise and music. The following figures show the performance of the coders for each test item (pooled for both test sites):

Figure 4: Codec-by-codec results for Programme Item 2 (English male voice, both test sites)

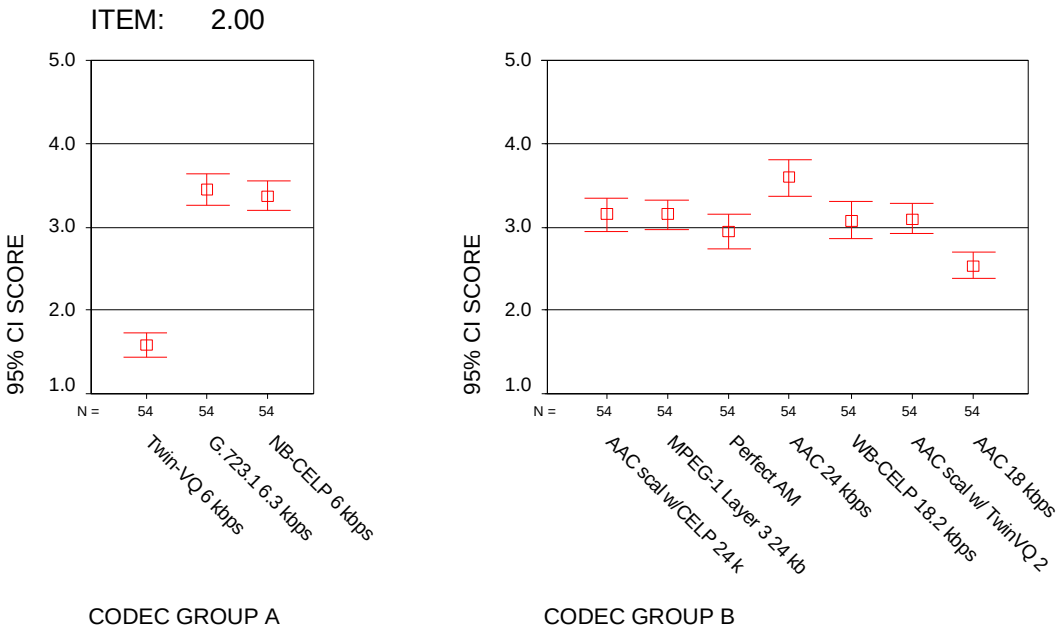


Figure 5: Codec-by-codec results for Programme Item 10 (English female voice, both test sites)

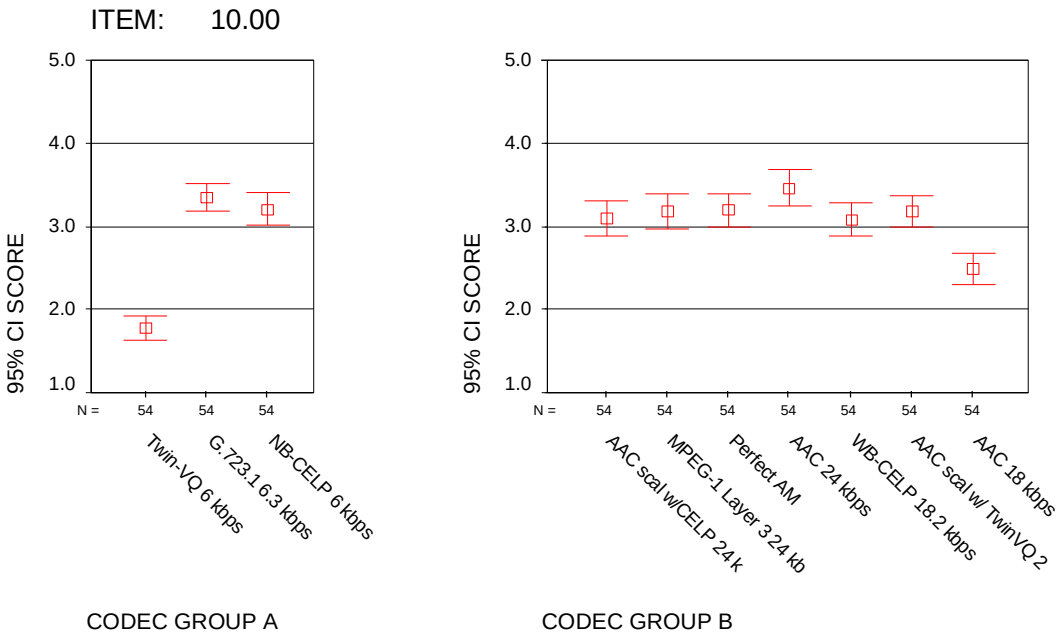


Figure 6: Codec-by-codec results for Programme Item 13 (French female voice, CCETT only)

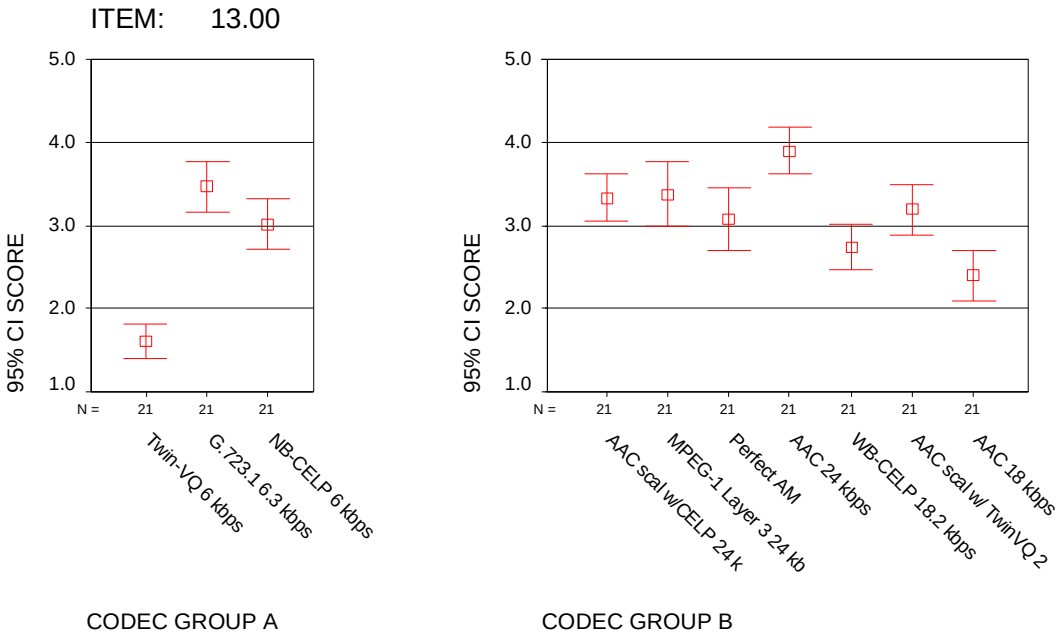


Figure 7: Codec-by-codec results for Programme item 20 (Speech + music, both test sites)

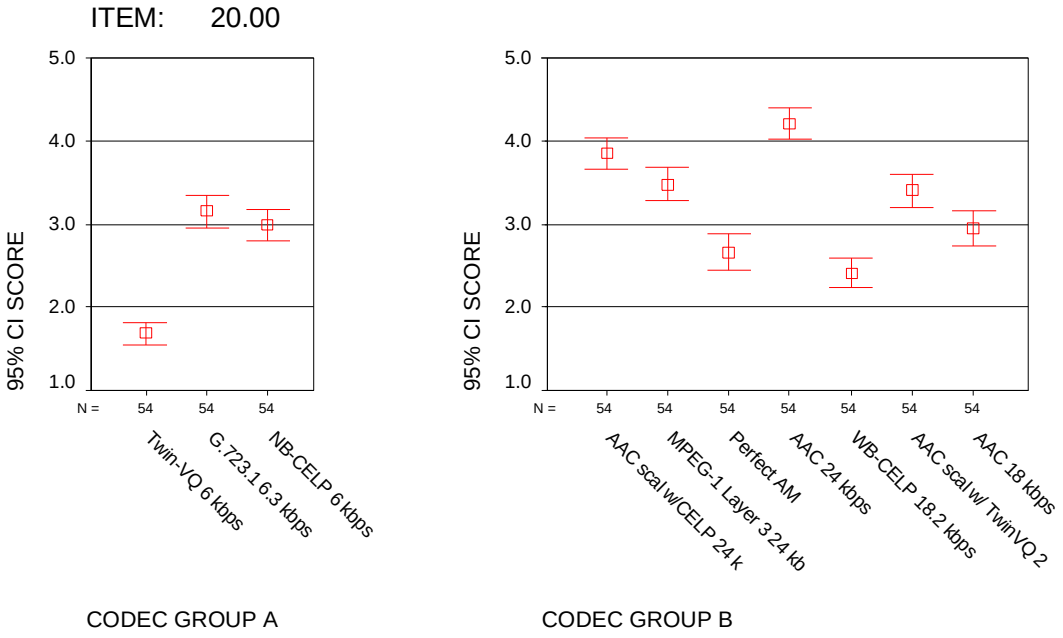


Figure 8: Codec-by-codec results for Programme Item 21 (Pop music, both test sites)

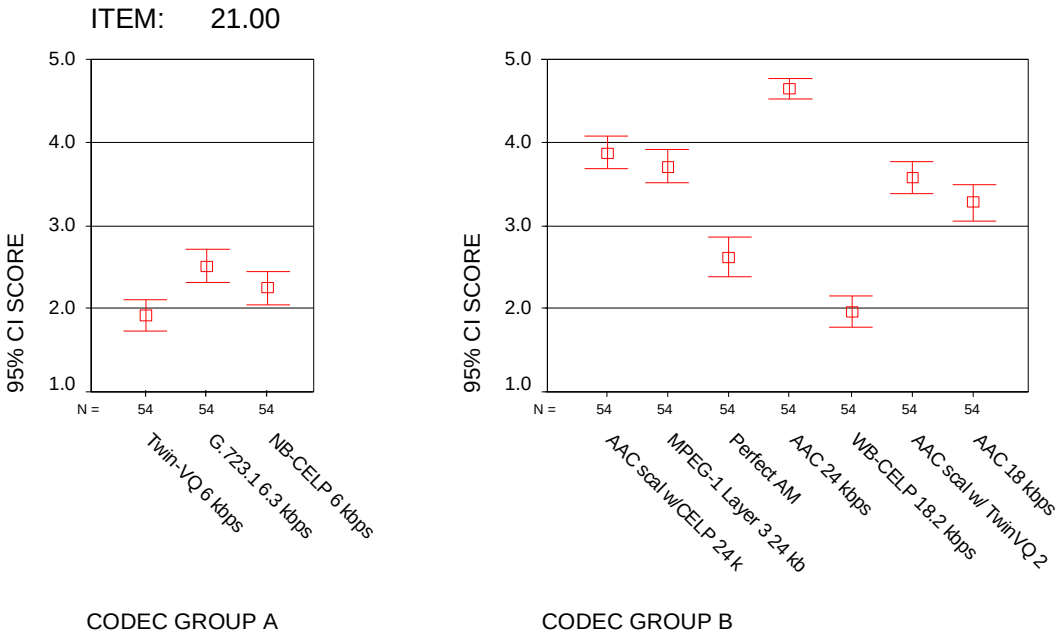


Figure 9: Codec-by-codec results for Programme Item 22 (Folk music, both test sites)

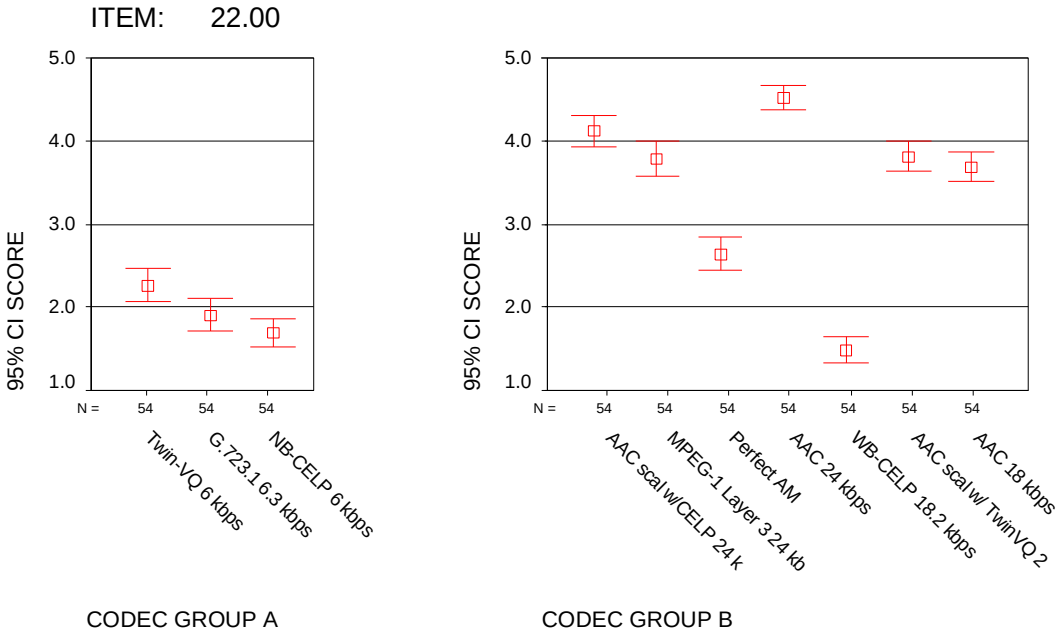


Figure 10: Codec-by-codec results for Programme Item 28 (French male voice + music, CCETT only)

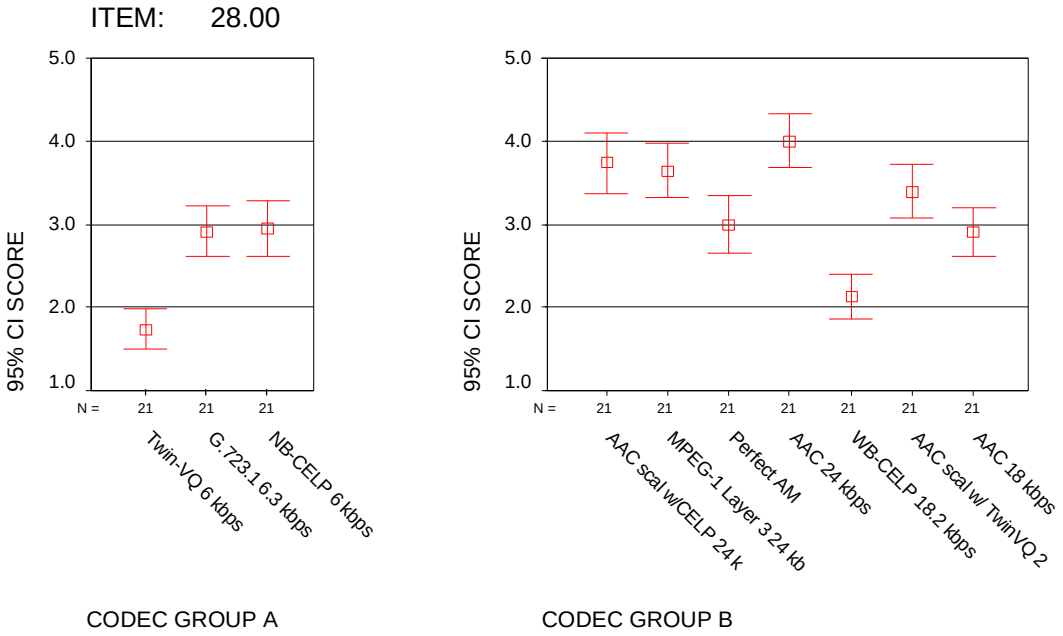


Figure 11: Codec-by-codec results for Programme Item 35 (Music+noise, both test sites)

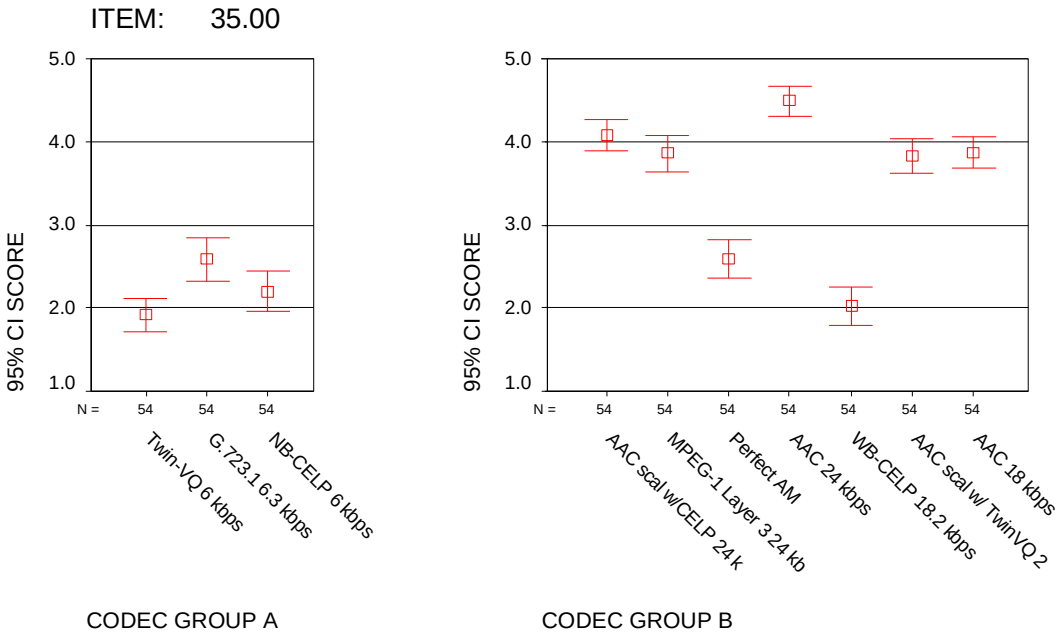


Figure 12: Codec-by-codec results for Programme Item 36 (Suzanne Vega, Teracom only)

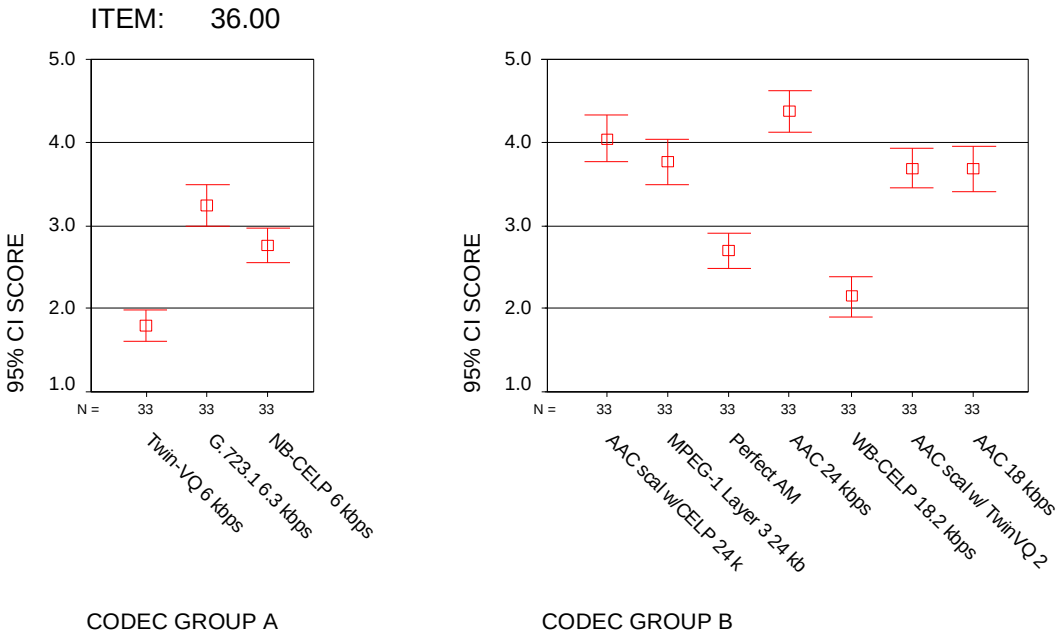


Figure 13: Codec-by-codec results for Programme Item 38 (Classic music, both test sites)

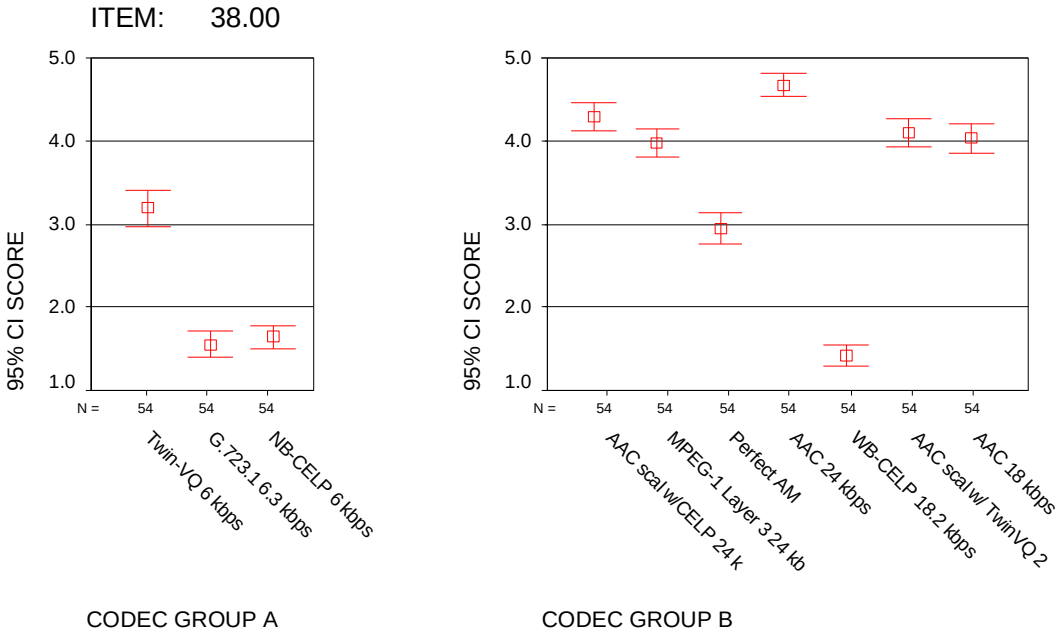


Figure 14: Codec-by-codec results for Programme Item 44 (Speech + noise, both test sites)

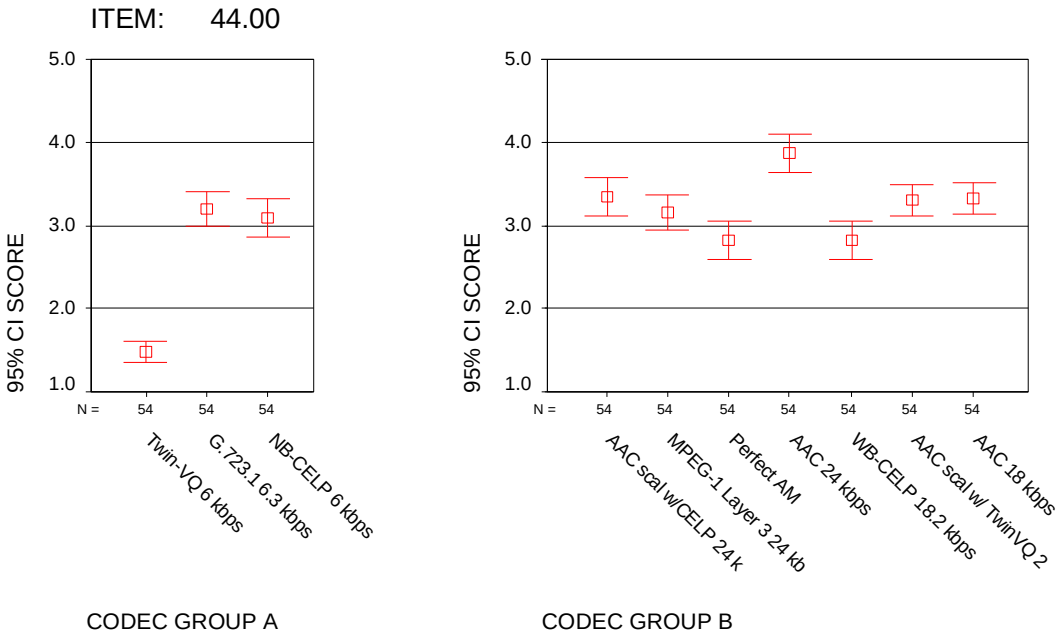
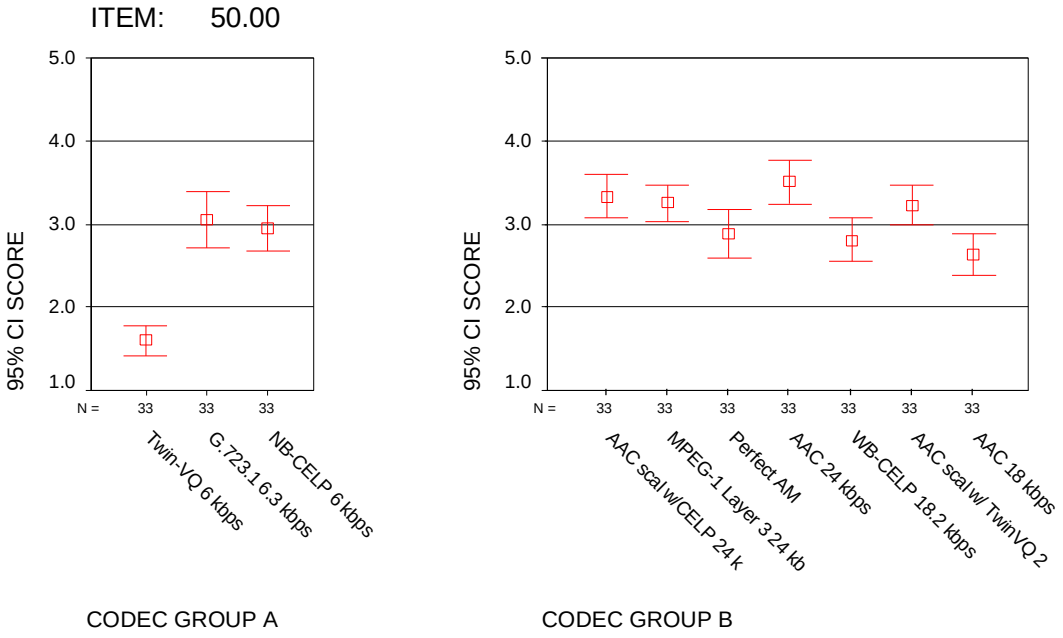


Figure 15: Codec-by-codec results for Programme Item 50 (Speech + noise, Teracom only)



9.5.4 Variance by programme items

«How does the performance of the codecs vary with programme item?»

In the 8 kHz test:

- TwinVQ shows relatively little variation by item; all items except #38 have mean rating between 1.5 and 2.5.
- G.723 varies a lot by item, with items ranging from mean score of 1.5 (#38) as high as 3.5 (#13).
- For NB-CELP, most items have mean rating between 2.5 and 3.5, but four items (#21, 22, 35, 38) have mean rating between 1.5 and 2.5

In the 24 kHz test:

- AAC/CELP was relatively consistent, with all items receiving mean rating between 3.0 and 4.5.
- MPEG-2 Layer III was very consistent, with all items receiving mean rating between 3.0 and 4.0.
- Perfect AM was the most consistent coder. All items received mean rating between 2.5 and 3.25, and only one item (#10) was coded significantly better than any other item (it received better ratings than #20, 21, 22, 35, and 36).
- AAC-24 was relatively consistent. Mean ratings ranged between 3.5 and 4.75, and there were significant differences in quality from item to item.
- The performance of the WB-CELP depends on the material to be encoded. Mean ratings ranged between 1.25 for the clean music items and 3.25 for the clean speech items.
- AAC/TwinVQ was relatively consistent, with all items receiving mean rating between 3.0 and 4.25.
- AAC-18 was the least consistent coder. Mean ratings ranged from below 2.5 to above 4.0.

Figure 16: Item-by-item results for Codec A1 (Twin-VQ 6 kbps)

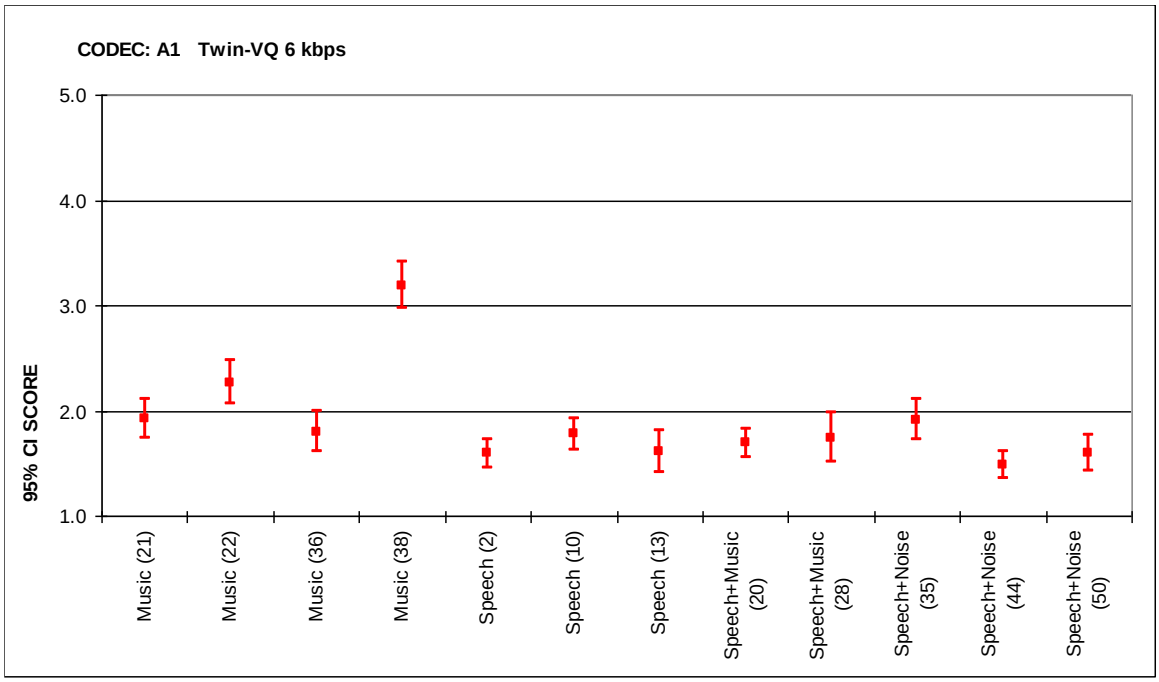


Figure 17: Item-by-item results for Codec A2 (G.723.1 6.3 kbps)

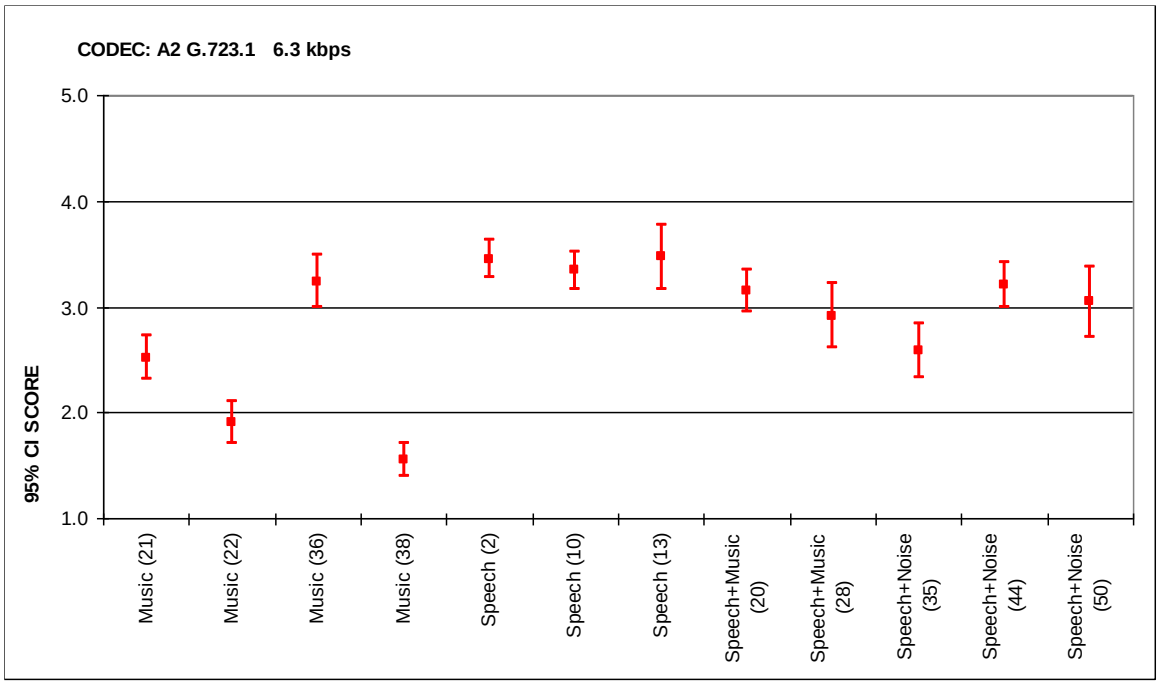


Figure 18: Item-by-item results for Codec A3 (NB-CELP 6 kbps)

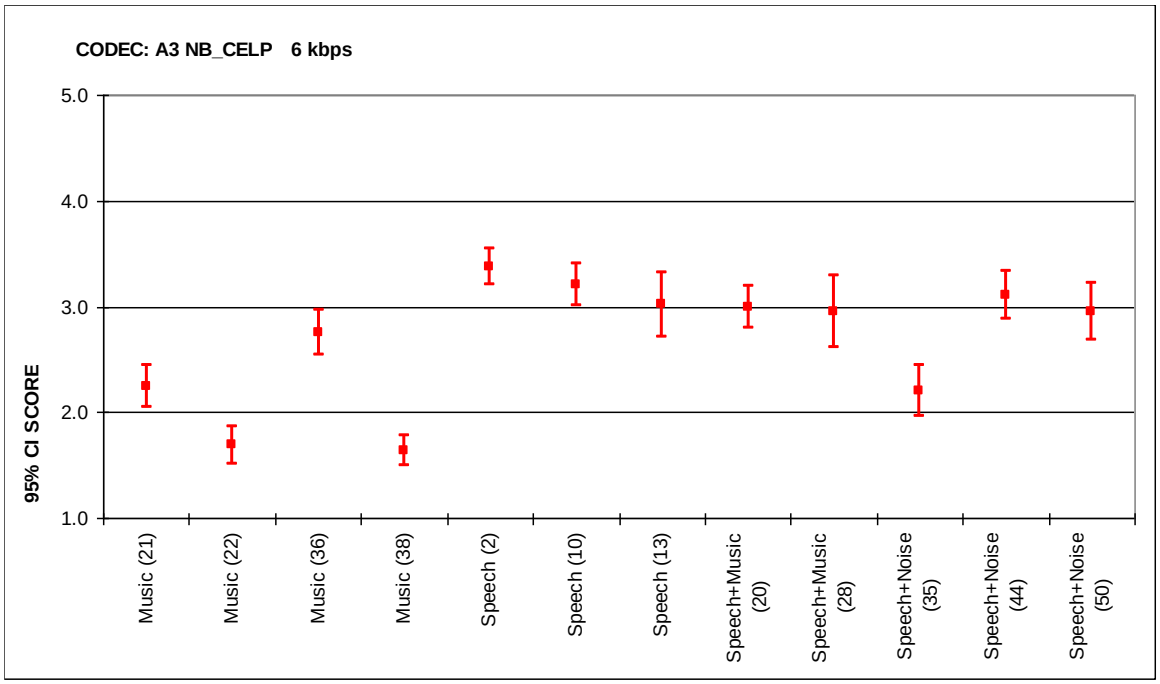


Figure 19: Item-by-item results for Codec B1 (AAC scal w/CELP 24 kbps)

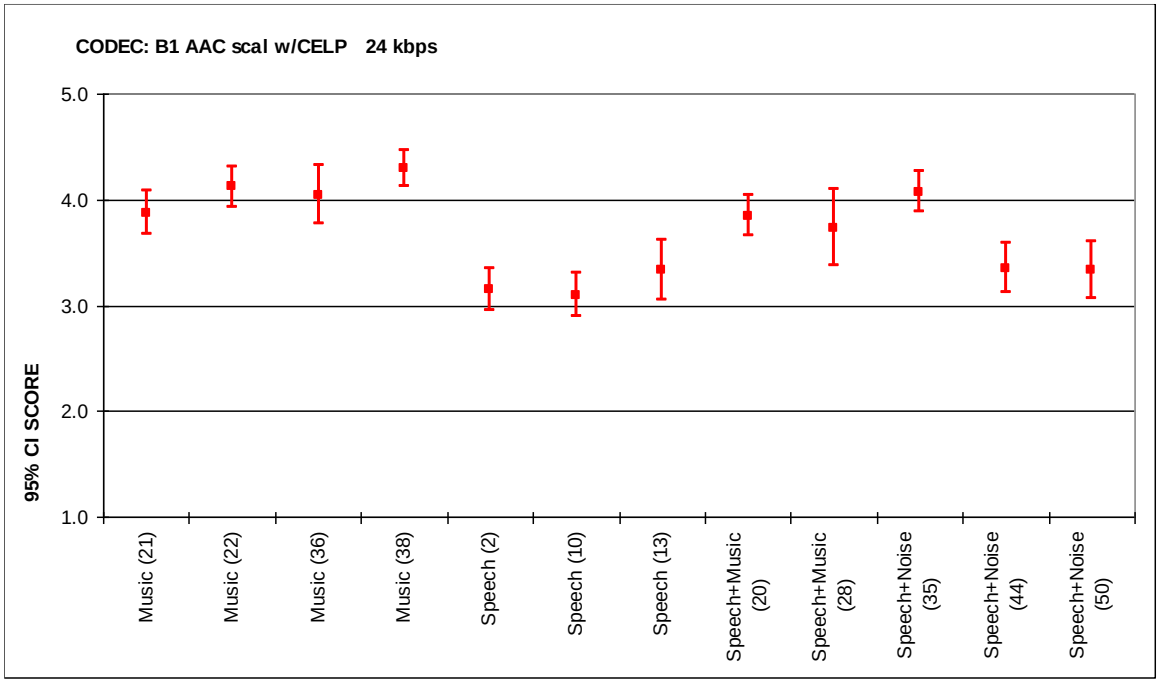


Figure 20: Item-by-item results for Codec B2 (MPEG-1 Layer 3 24 kbps)

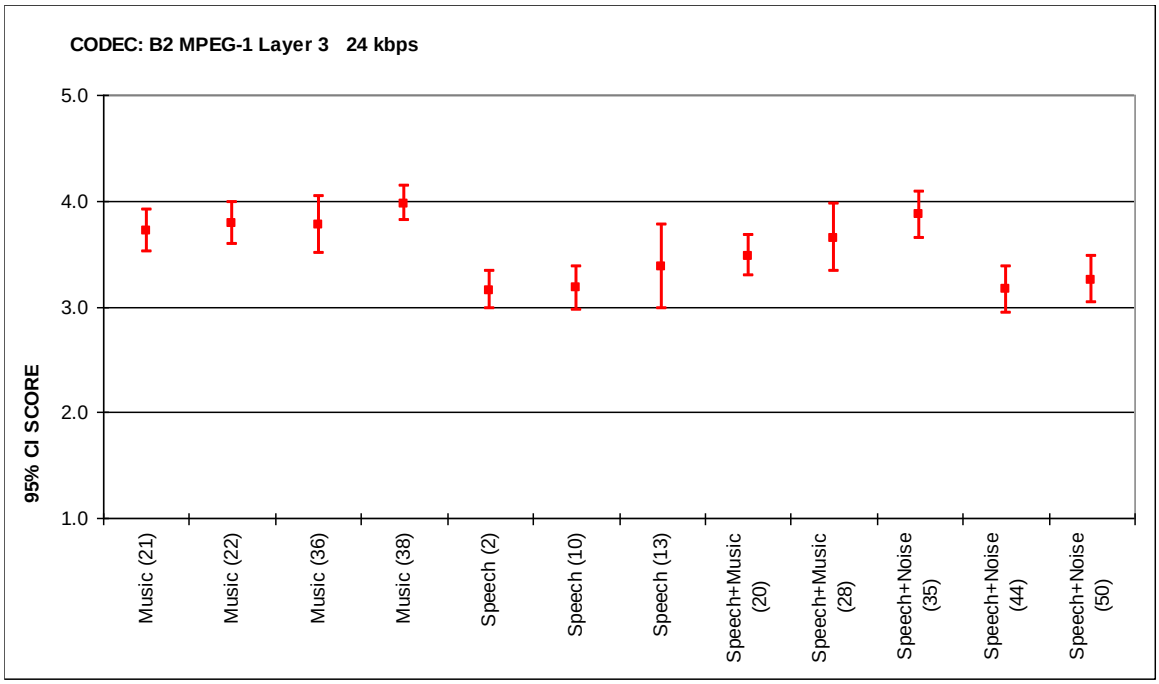


Figure 21: Item-by-item results for Codec B3 (Perfect AM)

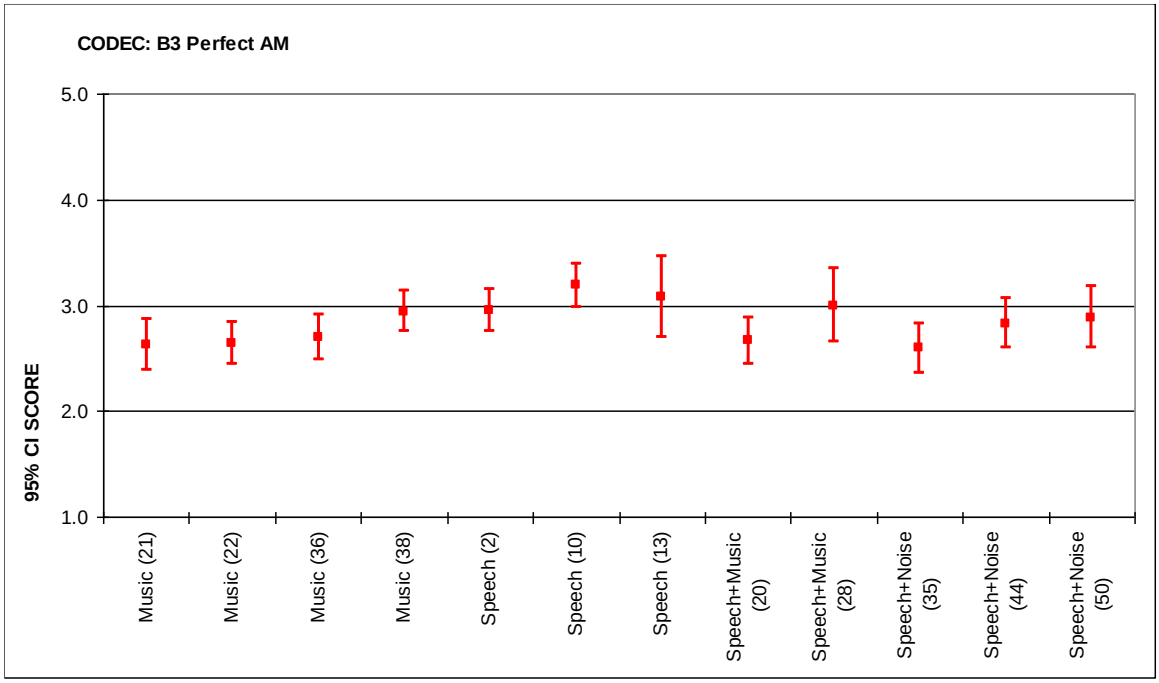


Figure 22: Item-by-item results for Codec B4 (AAC 24 kbps)

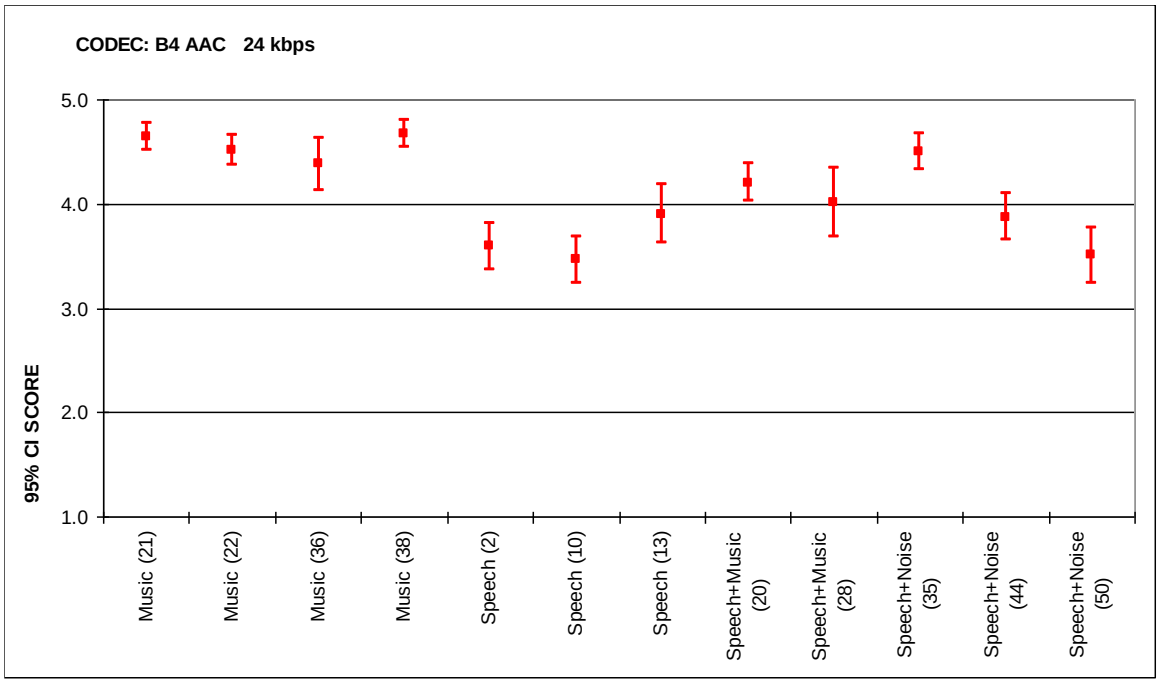


Figure 23: Item-by-item results for Codec B5 (WB-CELP 18.2 kbps)

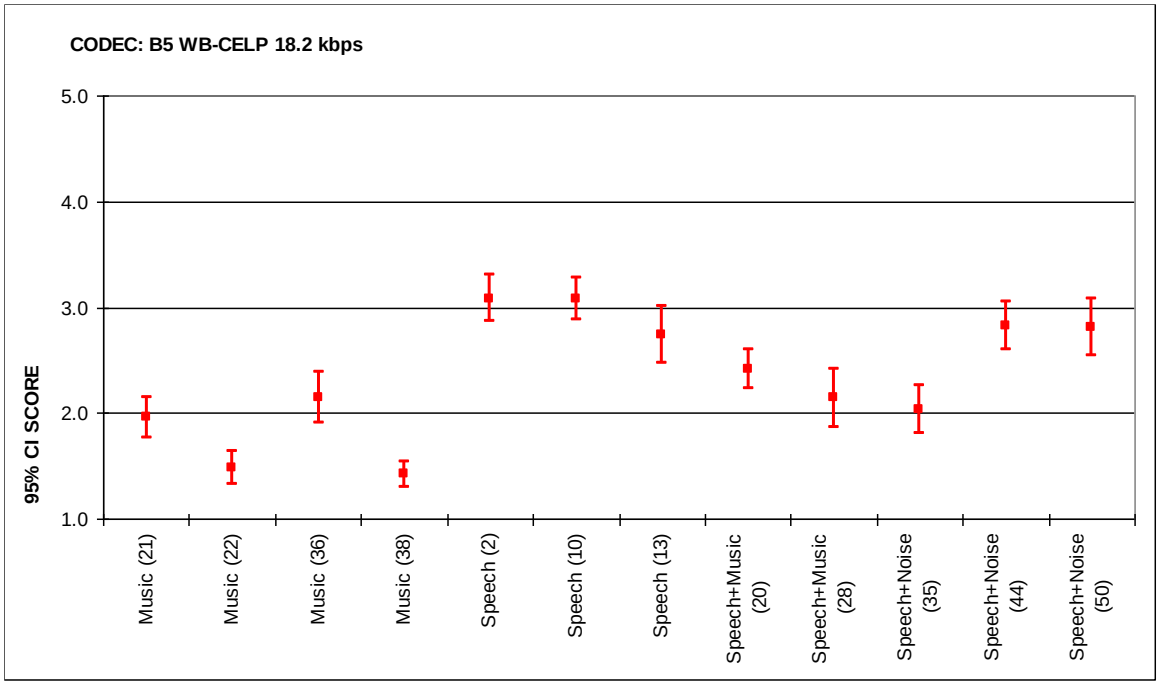


Figure 24: Item-by-item results for Codec B6 (AAC scal w/Twin-VQ 24 kbps)

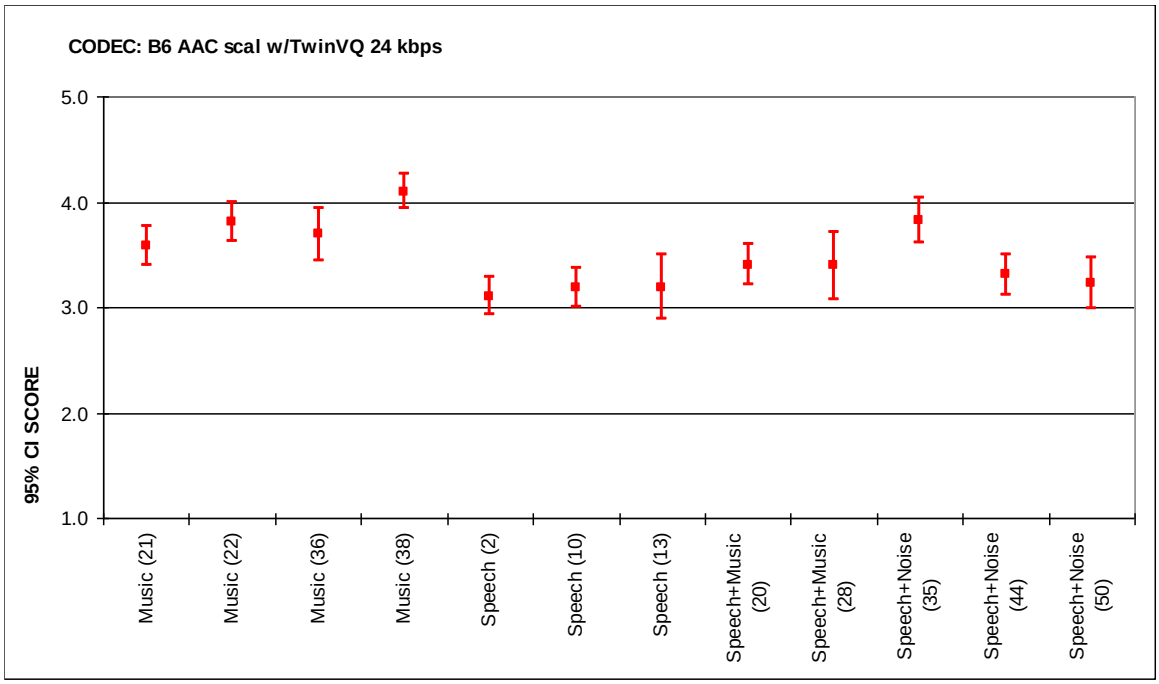
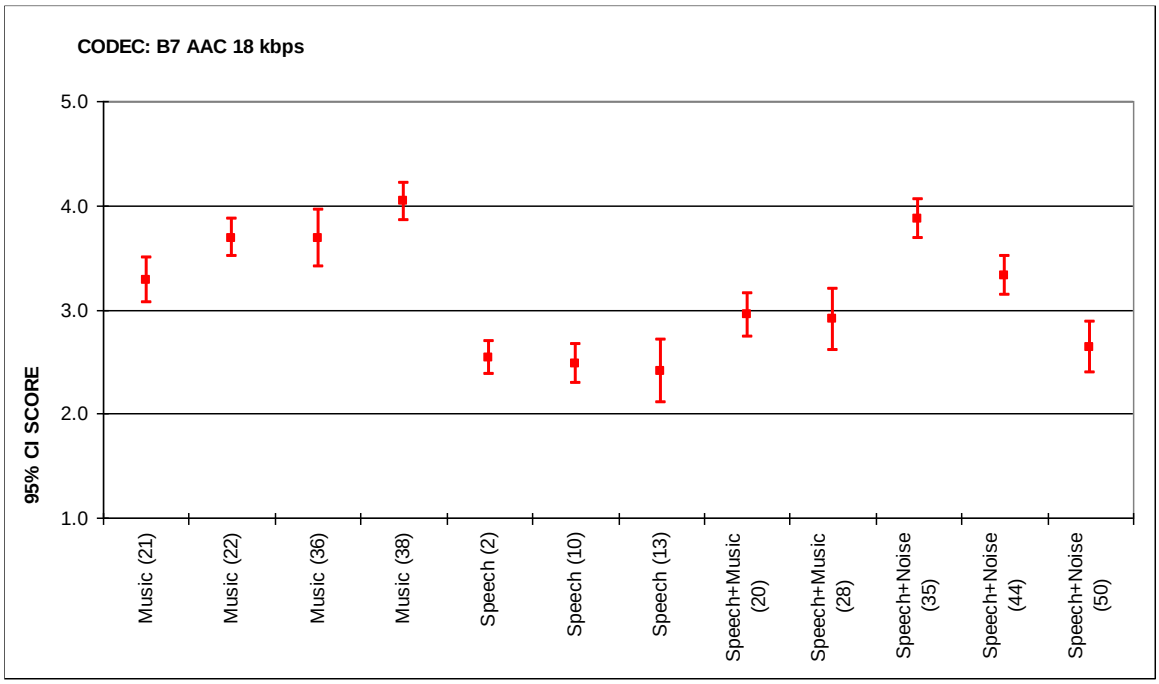


Figure 25: Item-by-item results for Codec B7 (AAC 18 kbps)



9.5.5 Ranking of codecs

«What is the relative ranking of the codecs tested?»

Taking both the overall ratings, and the pairwise comparisons into account, the codecs may be rated as follows.

In the 8 kHz test the NB-CELP test results are statistically equivalent to those of ITU-T G723.1 at CCETT or slightly worse at Teracom (although the confidence intervals of both coders overlap at Teracom, the coders are statistically different as a result of the Student t-test analysis). It has to be noted that the NB-CELP coder operates at a lower bitrate and has a shorter delay (see chapter 4). Both of these codecs performed better than Twin-VQ.

In the 24 kHz test, AAC-24 was the best codec, followed by AAC/CELP, MPEG-2 Layer 3 and AAC/TwinVQ. The ranking of the last three coders depends on the test site. . AAC/CELP and MPEG-2 Layer III were statistically equivalent at CCETT, MPEG-2 Layer III and AAC/TwinVQ were statistically equivalent at both test sites These four codecs were all significantly better than AAC-18, which was significantly better than AM, which was significantly better than WB-CELP.

9.5.6 1-layer vs 2-layer coding

«How is the performance of 1-layer AAC coding compared to scaleable (2-layer) coding at the same total bitrate?»

The AAC-24 coder performed significantly better than either of the scaleable codecs, at both test sites. Breaking down item-by-item, there were 7 of 12 items on which AAC-24 was superior to AAC/CELP and 9 of 12 for AAC/TwinVQ. There were no programme items for which either scaleable codec was superior to AAC-24.

AAC-24 versus AAC/CELP												
X = superior	Music				Speech			Speech+ Music		Speech+Noise		
	21	22	36	38	2	10	13	20	28	35	44	50
AAC-24	X	X		X	X		X			X	X	
AAC/CELP												

AAC-24 versus AAC/TwinVQ												
X = superior	Music				Speech			Speech+ Music		Speech+Noise		
	21	22	36	38	2	10	13	20	28	35	44	50
AAC-24	X	X	X	X	X		X	X		X	X	
AAC/TwinVQ												

Comparing AAC-24 to perfect AM, AAC_24 performed superior for 11 of 12 items, there was no item for which perfect AM was superior.

AAC-24 versus analogue												
X = superior	Music				Speech			Speech+ Music		Speech+Noise		
	21	22	36	38	2	10	13	20	28	35	44	50
AAC 24	X	X	X	X	X		X	X	X	X	X	X
Perfect AM												

The MPEG-2 Layer III codec overall performed as well as the AAC/TwinVQ scaleable codec at both test sites; it performed equivalently to AAC/CELP at CCETT, but AAC/CELP performed better at Teracom. Breaking down item-by-item, there were no items on which MPEG-2 Layer III performed differently than either scaleable coder.

9.5.7 Scaleable vs. multicast

«How is the performance of scaleable coding compared to multicast (base layer only) coding?»

The two scaleable coders, AAC/CELP, AAC/TwinVQ, both performed better than both multicast coders at both test sites. There were 9 items on which AAC/CELP was superior to WB-CELP and 8 on which it was superior to AAC-18; there were 8 items on which AAC/TwinVQ was superior to WB-CELP and 5 on which it was superior to AAC-18. There were no items in which either multicast coder was superior to either scaleable coder.

AAC/CELP versus WB-CELP												
X = superior	Music				Speech			Speech+ Music		Speech+Noise		
	21	22	36	38	2	10	13	20	28	35	44	50
AAC/CELP	X	X	X	X			X	X	X	X	X	
WB-CELP												

AAC/CELP versus AAC-18												
X = superior	Music				Speech			Speech+ Music		Speech+Noise		
	21	22	36	38	2	10	13	20	28	35	44	50
AAC/CELP	X	X			X	X	X	X	X			X
AAC-18												

AAC/TwinVQ versus WB-CELP												
X = superior	Music				Speech			Speech+ Music		Speech+Noise		
	21	22	36	38	2	10	13	20	28	35	44	50
AAC/TwinVQ	X	X	X	X				X	X	X	X	
WB-CELP												

AAC/TwinVQ versus AAC-18												
X = superior	Music				Speech			Speech+ Music		Speech+Noise		
	21	22	36	38	2	10	13	20	28	35	44	50
AAC/TwinVQ					X	X	X	X				X
AAC-18												

9.5.8 Scaleable vs. analogue

«What is the performance of the scaleable codecs compared to perfect analogue AM transmission?»

The two scaleable codecs both performed better than perfect AM simulation at both test sites. The AAC/CELP codec was superior to perfect AM on 8 of 12 items, and the AAC/TwinVQ codec was superior to AM on 7 of 12 items.

AAC/CELP versus analogue												
X = superior	Music				Speech			Speech+ Music		Speech+Noise		
	21	22	36	38	2	10	13	20	28	35	44	50
AAC/CELP	X	X	X	X				X	X	X	X	
Perfect AM												

AAC/TwinVQ versus analogue												
X = superior	Music				Speech			Speech+ Music		Speech+Noise		
	21	22	36	38	2	10	13	20	28	35	44	50
AAC/TwinVQ	X	X	X	X				X		X	X	
Perfect AM												

9.5.9 WB-CELP vs AAC-18

«What is the performance of AAC coding compared to MPEG-4 WB CELP, both at 18 kbps?»

Overall, AAC-18 performed better than WB-CELP at both test sites. On 8 of 12 items, AAC-18 performed better. These 8 items were speech+music, speech+noise and pure music. On 2 of 12 items, WB-CELP performed better. These 2 items were clean speech items (out of 3 clean speech items within the test).

AAC 18 mono versus WB-CELP												
X = superior	Music				Speech			Speech+ Music		Speech+Noise		
	21	22	36	38	2	10	13	20	28	35	44	50
AAC 18 mono	X	X	X	X				X	X	X	X	
WB-CELP					X	X						

10 Conclusions

The methodology and the statistical analysis of the results of formal listening tests that were conducted for the NADIB portion of the MPEG-4 audio verification testing have been presented. Several audio and speech coders have been tested. Please note that the speech coders were not designed for music which is present in several items used in this test.

The results of the analysis permitted the questions posed in the NADIB test plan to be answered and some general conclusions to be drawn for these tests.

- digital Audio in AM Bands using 24 kbit/sec audio coding has the potential to offer better quality compared to existing Analogue Modulation techniques,
- one layer encoding provides the best performance for a given net bit rate,
- if two levels of quality of service are targeted, one for normal reception conditions and one for bad reception conditions, a layered scaleable codec offers better quality than simulcast,
- in the scaleable mode, AAC+NBCELP is the best choice when programmes contain speech.
- with the encoders provided for this test, in the scaleable mode, when programmes contain music, TwinVQ performs better than NB-CELP but AAC+NBCELP performs somewhat better than AAC+TwinVQ.
- in the scaleable mode, AAC+NBCELP appears to be the best compromise for generic audio broadcasting amongst the coders in this test.
- on the material used in this test Twin VQ at 6 kbps performed worse than G.723.1 and NB CELP with the exception of some music items. TwinVQ is known to have lower quality for speech signals and also for speech+music, probably because speech signals are dominant when the bandwidth is limited below 4 kHz. Note that the TwinVQ under test directly quantizes input signals sampled at 24 kHz for the purpose of simple connection to AAC scaleable system. According to previous experience, it can be assumed that TwinVQ will achieve higher quality at the same bitrate if a lower sampling rate is used in the unscaled mode.
- the performance of the WB-CELP codec depends on the material to be encoded. For the speech items the results of WB-CELP and perfect AM are statistically equivalent (Note: this was not computed but concluded from the figures), for music items Perfect AM performs better.
- the NB-CELP test results are statistically equivalent to those of ITU-T G723.1 at CCETT or slightly worse at Teracom (although the confidence intervals of both coders overlap at Teracom, the coders are statistically different as a result of the Student t-test analysis). It has to be noted that the NB-CELP coder operates at a lower bitrate and has a shorter delay (see chapter 4).

11 Acknowledgement

The authors would like to thank all the persons that made these listening tests possible :

- the MPEG-Audio Test AHG members
- the NADIB CODEC group members,
- the selection panel members,
- MM JY Leseure , JP Thomas and H Maudier (CCETT) for their reliable support and help in test organisation at CCETT
- Personnel at the selection site (Swedish Broadcasting Corporation)
- Personnel at the test sites (CCETT and Teracom)

The authors would like also to thank Mrs Laura Contin (CSELT) and MM Pete Schreiner (Scientific Atlanta), David Meares (BBC), Martin Dietz (FhG) and Jean-Bernard Rault (CCETT) for their valuable contribution to this document.

12 Bibliography

[1] ISO-IEC/JTC1/SC29/WG11/N1849, “Proposal of MPEG-4 Audio verification tests", October 1997
[2] ISO/IEC JTC1/SC29/WG11/m2816, “Proposal for Evaluation of MPEG-4 Audio Technology for the NADIB Application", October 1997
[3] ISO/IEC JTC1/SC29/WG11/m3047, “Proposal of NADIB verification tests", February 1998
[4] ISO/IEC JTC1/SC29/WG11/ N1729, “MPEG-4 Applications Document", July 1997
[5] ISO/IEC JTC1/SC29/WG11/N2157, “MPEG-4-Audio verification test specification", March 1998
[6] ITU-R, "Subjective Assessment Of Sound Quality, Recommendation 562-3", 1990

13 ANNEXES

13.1 ANNEX 1: Test Schedule

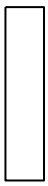
Activity	Deadline	Responsibility	Comments
Test material on ftp-site	28 February 98	Deutsche Welle, Sony, CCETT, Swedish Radio, Berkom	
Pre selection & Pre processing	9 March 98	Samsung, FhG, Deutsche Welle, Teracom	
Resampling, delivery to the ftp site	23 March 98	Samsung, University of Hannover	generate 8,16 & 24 kHz originals
Coding process including the AM version	1 April 98	FhG, NTT, NEC, Philips, Deutsche Welle, CCETT	
Decoding, Upsampling CD-ROM to Swedish Radio & other sites	6 April 98	FhG	must include bandlimited reference (fs=24kHz)
Test Equipment set-up	9 April 98	Swedish Radio	
Bitstream/bitrate & decoding verification	17 April 98	NTT, CCETT, CSELT, NTT, Philips, Samsung	
Selection process	17 April 98	Swedish Radio	L.Mossberg(Swedish Radio), W. Schäfer (Sony) J.-Y. Leseure (CCETT) N. Schall (Deutsche Welle)
Aural announcements on ftp site	17 April 98	NTT	
Test preparation	27 April 98	Teracom, CCETT, Swedish Radio	Computer based test set-up Visual announcement 4 randomisations
Grading phase	22 May 98	Teracom, CCETT	
Statistical analysis	19 June 98	MIT	
Test report	26 June 98	CCETT, MIT, Teracom	

13.2 ANNEX 2: Test organisation at Teracom

13.2.1 Test procedure

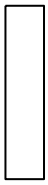
For each trial both the reference, A, and the processed version, B, were presented as A-B-A-B. Between each trial an aural announcement indicated the actual number of the trail (”item nn”). This information was also presented on a computer screen together with an indication whether the reference or the processed version was being played (for example: 12 B). After the A-B-A-B presentation there were eight seconds of silence during which the listeners gave their judgements.

As the CCETT decided to use the French version of the continuous quality scale, it was logical to use a Swedish version at Teracom. The scale is shown below:



5.0 Mycket Bra
4.0 Bra
3.0 Varken Bra eller Dålig
2.0 Dålig
1.0 Mycket Dålig

The English counterpart is given below:



5.0 Excellent
4.0 Good
3.0 Fair
2.0 Poor
1.0 Bad

The grades were written on voting sheets, one for each session, carefully labelled with randomisation number, session number, date, name, age, profession and expert/non expert. A voting sheet is annexed to this document, see **ANNEX 3**.

13.2.2 Listening conditions

The tests were carried out in a recently built and specially designed listening room. The room is expected to, in principle, fulfil the requirements given in the ITU-R Recommendation BS.1116. Stax Lambda Pro headphones were used during the test. Groups of up to five listeners at a time could listen simultaneously. The diffuse field equaliser was turned off and the listening level was set to 21 on the headphone driver. These conditions were the same for both test sites.

Each listening session had a duration of approximately 25 minutes and was followed by a break. The total test, including breaks, lasted six hours. The test was divided in two parts where the first part (the A-test) included low bit rates (6 kbit/s) speech codecs [5] and the second part (the B-test) included higher bit rates (16 kbit/s and 24 kbit/s). Each part was initiated by a training session where the listener became familiar with the quality range of the test and the test procedure. The A-test was divided into two listening sessions whilst the B-test consisted of five sessions. The time schedule is found below.

Activity	Time

Training A-test	30 minutes
Break	5 minutes
Session A.1	25 minutes
Break	10 minutes
Session A.2	25 minutes
Break	20 minutes
Training B-test	20 minutes
Break	5 minutes
Session B.1	25 minutes
Break	10 minutes
Session B.2	25 minutes
Lunch	60 minutes
Session B.3	25 minutes
Break	10 minutes
Session B.4	25 minutes
Break	20 minutes
Session B.5	25 minutes

13.2.3 Training

The training was carried out according to the following procedure:

- Introductory discussion, where the purpose of the test was explained and the listeners read the listener instructions included at the end of this document (Annex 3).
- The test procedure and the quality scale were discussed. It was specially pointed out that the scale was continuous and that impairment types could be discussed, however, the grades should absolutely not be discussed with the other listeners. It was emphasised that it was the individual's judgement that was being sought. It was also mentioned that the reference should be considered as an indication of the intended quality for each program item, i.e. it should correspond to an excellent quality.
- The group listened to each of the training items and discussed which impairments they noticed. Simultaneously, the listeners were invited privately to grade the training items and thereby got used to the test procedure.

13.2.4 Listeners

36 listeners (9 women and 27 men, aged between 20 and 60 years old) participated in the test. 25 of them were non experts while 11 were skilled listeners.

13.2.5 Verification of results

The two following tables display the scoring differences for the ten duplicated samples of the subjects at Teracom. The analysis and interpretation of these data is presented later (Section 13.3).

Item/codec	2/A3	2/A3	36/A2	36/A2	44/A1	44/A1
grading	Average	diff.	Average	diff.	Average	diff.
Subject 1	3,75	0,3	3,75	-0,3	1,8	0,6
Subject 2	2,25	-0,5	2,5	0	1,25	-0,5
Subject 3	2,5	0	2,5	0	1	0
Subject 4	3,6	-0,2	3,65	-0,3	1,9	-0,2
Subject 5	2,85	0,1	2,45	0,1	1	0
Subject 6	3,15	1,3	3,6	-0,8	1,75	-0,1

Subject 7	3,85	0,3	4,2	0,4	1,7	1,2
Subject 8	3,95	0,1	2,15	-0,1	1	0
Subject 9	2,5	0,6	2,5	0,6	1,25	0,5
Subject 10	3,1	0,8	1,95	-0,1	1	0
Subject 11	3,25	-0,5	2,65	-0,7	1,85	-0,7
Subject 12	3,85	-0,3	3,1	1,4	1,65	0,3
Subject 13	2,9	-0,2	2,25	-0,1	2	0
Subject 14	4,25	0,5	4,45	0,9	1,75	-0,5
Subject 15	4,15	-0,1	4	0,4	1,65	-0,9
Subject 16	3,85	0,3	4,15	0,3	1,55	-0,5
Subject 17	3,5	-0.5	2,7	0,2	1	0
Subject 18	4,25	-0,5	4	0	2,25	-0,5
Subject 19	3,95	0,1	4,5	-1	3,1	-0,2
Subject 20	3,8	0,2	2,9	-0,4	1,85	0,1
Subject 21	3	-1	3,2	0,6	1,55	-0,1
Subject 22	3,15	-0,7	3,75	0,5	1,5	0
Subject 23	3	0	3,75	0,5	3,5	0
Subject 24	3,25	-0,5	3,75	-0,5	2,35	0,3
Subject 25	3,4	-0,8	3,75	-1,5	2,8	0
Subject 26	2,9	0	3,9	0	1,65	0,3
Subject 27	3,8	-0,2	3,05	-0,1	2,05	0,9
Subject 28	3,65	-0,3	3,3	-0,6	1,55	0,5
Subject 29	4,4	0,8	3,5	-1	1,5	0
Subject 30	4,15	-0,3	4,35	-0,1	1,65	-0,3
Subject 31	2,85	0,1	2,3	0,6	1,15	0,1
Subject 32	2,85	0,1	2,95	-0,1	1,3	0
Subject 33	3,55	-0,5	3,25	0,1	1,6	-0,8
Subject 34	4,05	0,3	3,55	0,3	2	0
Subject 35	3	1,2	3	0	1,5	0
Subject 36	3,25	0,5	3,65	-0,3	1,2	0

Item/codec	10/B5	10/B5	20/B2	20/B2	21/B6	21/B6	22/B7	22/B7	35/B3	35/B3	38/B1	38/B1	50/B4	50/B4
grading	Aver.	diff.	Aver.	diff.	Aver.	diff.	Aver.	diff.	Aver.	diff.	Aver.	diff.	Aver.	diff.
Subject 1	2,9	0,4	3,05	0,5	3,4	0,8	3,5	1	2,25	1,5	3,65	1,5	2,55	0,1
Subject 2	2,5	0	2,5	0	4	0	3,5	-1	2,25	-0,5	4,5	0	3,5	1
Subject 3	3,45	0,1	3,3	-0,2	2,85	-0,5	3,25	0,1	2,45	0,7	3,95	0,1	4,15	-0,7
Subject 4	3,65	-0,3	4,35	-0,5	4,25	-0,5	4,65	-0,3	3,8	0,4	4,35	-0,3	4,2	0,6
Subject 5	2,9	-0,2	2,7	-0,2	3	0	4,15	-0,3	3,4	-1	3,65	-0,7	3,25	-0,5
Subject 6	2,65	-0,3	3,35	-0,7	3,4	-0,4	4,15	0,3	3,4	-0,2	3,95	-0,7	3,1	-0,4
Subject 7	2,45	0,7	3,8	-0,8	3,75	-1,1	3,8	0,4	1,35	0,1	4,9	0	2,5	0,4
Subject 8	4,1	-0,8	3,2	0,6	2,75	-0,5	2,65	-0,3	1	0	4,9	0,2	2,65	-0,3
Subject 9	2,3	-0,4	2,5	-0,6	3	1,6	3,75	-0,5	1,75	-0,5	3,25	-0,5	2,25	-1,5
Subject 10	1,3	0,6	2,5	-1	3	0	2,85	-0,1	1,2	0	3,45	0,1	2,7	0,2
Subject 11	2,7	-0,4	3,75	-0,5	3,85	0,1	3,8	0,2	2,35	0,3	3,8	0	3	0
Subject 12	3,05	0,1	3	-0,6	3,3	0,2	3,35	0,7	2,85	-0,3	4,15	-1,3	2,8	-0,4
Subject 13	2,25	0,3	3	0	3,25	0,1	3,5	1	2,65	-0,3	3,6	0,2	3,1	0,2
Subject 14	2,8	0	3,1	1,8	4	1	3,65	0,3	2,65	0,7	4,95	-0,1	3,95	1,9
Subject 15	3,35	0,9	3,85	0,9	3	0,4	3,85	-0,5	2,35	0,5	4,8	0	3,9	-0,4
Subject 16	3,85	0,1	3,4	0,8	3,9	0	4,1	-0,4	2,85	0,3	4,35	-0,3	3,8	-0,2
Subject 17	2,9	0,2	3,4	0,8	3,9	0,2	3,85	-0,7	2	1	4,75	-0,1	4	-0,4
Subject 18	2,75	0,5	4,35	0,3	4,25	0,5	3,5	0	2	1	4,65	-0,3	4,5	0
Subject 19	3,35	0,7	3,4	-0,8	3,25	0,5	4	0	3,4	1,2	4,8	0	4,35	-0,7
Subject 20	2,75	-0,3	3,65	-0,3	3	0,2	3,3	1,2	2	0,2	3,55	-1,3	3,8	-0,2
Subject 21	2,5	0	3,05	0,1	3,2	0,6	3,6	1,2	2,45	0,9	4,75	-0,1	4,2	0,4
Subject 22	1,95	-0,1	3,9	1,6	4,25	0,9	4,7	0	3,15	0,7	4,45	-1,1	4,35	-0,9
Subject 23	2,25	0,5	3,5	0	3	0	3,25	0,5	2	1	4	-1	3,5	0
Subject 24	3,6	0,2	3,35	-1,3	3,6	-0,8	3,85	-0,3	3,5	0	4,35	-0,3	4,1	-0,2
Subject 25	1,6	0,8	2,75	0,5	3,7	-0,4	4,1	0,2	3,65	-0,3	4,1	-1,8	3,95	-1,7
Subject 26	3,2	0,4	3,35	-0,1	3,55	-0,5	3,65	0,9	2	0	4,3	0,4	3,7	0,2
Subject 27	3,1	0	3,2	0	3,3	-0,4	4	0,4	2,2	0,6	4,2	-0,4	4,4	0,2
Subject 28	3,35	0,3	3,8	-0,2	3,5	-0,2	4,35	-0,7	3,1	-0,2	4,8	0	3,55	-0,9
Subject 29	3,8	0	4,3	0,4	3,7	0,4	4,2	0,2	2,75	0,5	4,85	-0,1	4,55	0,5
Subject 30	3,7	1	4,35	-0,3	4,3	0	4,35	0,3	3,65	0,3	4,75	-0,1	4,5	0
Subject 31	2,8	0	3,05	0,5	3,15	0,3	4	0	3	0	4,45	0,5	2,8	-0,2
Subject 32	3,15	0,5	2,55	0,3	4,65	0,7	4,65	0,3	3,15	1,5	5	0	3	0,6
Subject 33	2	0	2,95	0,5	2,9	0	3,35	0,9	2,25	-0,3	4,35	0,3	2,7	-0,2
Subject 34	2,65	-0,9	3,7	0,8	4,35	-0,7	4,55	0,1	3,7	1,6	4,85	0,1	2,85	-0,1
Subject 35	3,1	-0,2	3,15	0,7	4,2	0	4,4	0,8	3,2	0,4	4,65	0,7	3,85	0,1
Subject 36	3,15	-0,3	3,4	1,2	3,85	0,1	3,25	0,5	2,45	-0,9	4,05	-0,5	4	0

13.3 ANNEX 3: Test organisation at CCETT

13.3.1 Listening conditions

The test has been carried out with two consecutive phases. The first phase was dedicated to the assessment of the narrow band codecs (A test), the second phase was dedicated to the assessment the wide band codecs (B tests). Within each phase, the test has been divided in sessions of 20 to 25 minutes length. A Test and B Test were preceded by a short training session. Between two consecutive sessions, listeners took a break of at least 15 minutes duration.
22 non-expert listeners performed the A-test and the same 22 plus one performed the B test. During the sessions, there were a maximum of four listeners at a time.
Each listener had two sessions per day. In consequence, the tests were completed over a 5 days period per listener.

13.3.2 Test equipment

The tests were run on STAX Lambda Pro headphones. The diffuse field equaliser was OFF. The loudness was set up empirically by the sound engineer of CCETT (JY Leseure), with informal subjective listening tests before the start of the training phase, at a level offering a comfortable listening level (for him). During the training sessions, the loudness level was discussed with the listeners and none of them complained with the proposed level, neither during the training phase nor the real test phase.
In order to give an indication of the headphone gain setting used, the white dot on the level adjustment screw faced the graduation 21 on the Stax Lambda headphones amplifiers.
The audio stimuli used for the training sessions and the grading sessions were digitally recorded onto optical disks using a SONY DD1000 recorder. This was done from a Silicon Graphics workstation where the items were first stored in a wave format. During the listening sessions the optical disk player was driven by the video player used for the visual announcements so that the sounds and pictures were synchronised.

13.3.3 Announcement

It has been decided to use mixed aural and visual announcements to help the listeners in the test progress.
The aural announcements concerned the session number the listeners were taking part in as well as the item number in the session they had to grade:
"session one" "item one" "item two" "item three"
The aural announcements were pre-recorded with the AB AB test stimuli on Optical Disks and it was decided to use a natural French voice.
The visual announcements were displayed on a TV screen, for each presentation, via a pre-recorded video tape and in accordance with the audio time code. The following sequence was displayed : "A", "B", “A", "B" and "VOTE".

13.3.4 Subjects

In order to fit the test recommendation, 22 non expert listeners participated in the listening sessions. They were divided in 5 groups of 4 listeners + 1 group of two listeners. Most of them are students (aged between 20 and 30), and there were more or less as many women as men.
There was no audiometric tests as the listeners were supposed to represent the average population in terms of their hearing capabilities.

Nevertheless, subjects reliability can be verified in a first stage by comparing their scoring on the repeated trials (shown in bold in **ANNEX 4**) and in a second stage by post-processing the results (done by the analysis centre).

13.3.5 Grading & Instructions for Scoring

As already written above, the grading scale that has been used is the BS 1284 5-grade scale, and it has been used as a continuous quality scale with one decimal place.

Each listeners had first to read a paper in his native language containing all the instructions. **ANNEX 3** contains the English version of those instructions, and its exact translation in French.

The listeners were invited to ask any questions they wanted in order to clarify everything.
During the training phase, they were advised to score the items so that they got used to using the quality scale. **ANNEX 3** contains the English and French versions of the scoring sheet. After the training phase, again the listeners were free to ask any questions or to do any suggestions to improve the quality of the test. Then, the formal "real" tests were run.

13.3.6 Training of the Subjects

In order to train the subjects prior to both tests A and B, one training session per test was performed. During that time, all the codecs were played (according to the test blocks shown in figure 1) with the items specially selected for the training phase. The purpose was to show the range of quality / artefacts that were present. There was a guidance and support before and after training from the test site personnel.
The listeners were given a general introduction to the tests. As with the main tests, each listener could score the training test individually although the results were discarded. It was only for the listener familiarisation to the test.

13.3.7 Verification of results

The results for these tests were recorded on paper by the listeners, together with careful labelling, including date, age, name, session number ... (see **ANNEX 3**) . Each set of results was entered into an excel sheet and immediately double checked to ensure no mis-entering had taken place.
As already mentioned, some trials were repeated in order to check the listeners scoring. Here are the scoring difference at CCETT :

item	**44**	**44**	**2**	**2**	**13**	**13**
codec	**1**	**1**	**3**	**3**	**2**	**2**
Grading	Average	Difference	Average	Difference	Average	Difference
listener_1	1.7	-0.2	4	-0.2	4.3	0
listener_2	1.7	-0.4	3.85	0.9	4.3	0
listener_3	1.45	-0.5	3.9	0.6	3.7	0.4
listener_4	1.5	0	3.6	0.2	3.8	-0.4
listener_5	1.1	-0.2	3	0	2.8	0
listener_6	1.2	-0.2	3.4	0.8	3.9	-0.4
listener_9	1.15	-0.1	2.6	-0.6	2.85	-0.3
listener_10	1.75	0.3	3.4	-1.2	3.85	-0.1
listener_11	1.75	-0.5	4.4	-0.4	4.7	0.2
listener_12	1.15	0.3	2.65	-0.5	2.6	0.4
listener_13	2.15	-1.3	4.2	0	3.9	-0.2
listener_14	1.25	-0.5	2.95	-0.1	2.7	-0.4
listener_15	1	0	2.8	0.4	2.8	0
listener_16	1.35	0.3	3.9	0.2	3.05	-0.1
listener_17	1.8	0	2.9	-0.2	2.65	0.3
listener_18	1	0	2.55	0.7	3.25	-0.5
listener_19	1.05	-0.1	2.75	-1.1	4.15	0.1
listener_20	1.3	0	3.5	0.6	2.65	0.7
listener_21	1.05	-0.1	2.7	-0.4	2.8	0.6
listener_22	1.1	0	3.6	-0.6	3.9	-0.4
listener_23	1.1	0	3.55	0.9	3.45	-0.1
listener_24	1.4	-0.8	4	-1	4.3	-0.2

Average and difference in scoring for each listener, and each repeated trial for A test.

item	284	284	227	227	105	105	353	353	381	381	202	202	216	216
codecB	4	4	7	7	5	5	3	3	1	1	2	2	6	6
Grading	Average	Difference	Average	Difference	Average	Difference	Average	Difference	Average	Difference	Average	Difference	Average	Difference
listener_1	3.85	0.3	3.15	-0.1	3.15	1.7	1.6	-0.4	4.1	-0.2	3.45	0.1	3.4	-0.8
listener_2	3.1	0.4	3.8	-1	2.85	0.1	1.95	0.1	4.65	-0.3	4.55	-0.1	4.25	-0.1
listener_3	3.55	1.3	3.35	-1.1	4.35	0.1	2.45	1.1	4.6	-0.4	3.85	0.9	3.75	0.7
listener_4	3.7	1.6	4.3	-0.2	2.7	0.2	2.55	-1.1	4.05	-0.5	2.7	0	4.45	0.5
listener_5	2.9	0	3.15	0.3	2.7	0	1.15	-0.3	4.25	0.5	3.3	0.8	2.75	0.3
listener_6	4.9	-0.2	3.95	0.5	2.45	0.3	2.9	0	5	0	4.4	-0.8	4.4	-0.8
listener_7	3.15	-0.5	3	-0.2	2	-0.6	2.45	-0.5	3.6	0	2.95	-0.1	3.05	-0.3
listener_9	4	-0.6	3.3	-0.4	3.55	0.9	1.65	-0.3	4.1	0.6	2.5	0	2.7	0
listener_10	3.45	-0.1	3.45	-0.5	2.75	-0.5	2.05	-0.3	4	-1.6	3.05	0.3	2.95	0.5
listener_11	4.7	0.4	4.25	0.1	3.3	1	3.55	0.7	4.7	0.2	4.05	0.1	3.95	0.9
listener_12	4.5	0.6	3.15	-2.1	3	0.6	2.1	-0.6	4.5	-0.4	2.65	0.5	3.2	0
listener_13	4.35	0.3	3.9	-0.2	4.1	-0.2	2.5	-0.6	3.85	0.7	3.35	0.3	3.5	0
listener_14	4.25	-0.5	4.25	-0.5	3.1	1.8	3.15	0.7	5	0	4.65	-0.3	3	0
listener_15	4.35	0.7	4.5	-0.6	4.05	1.1	3	0	4.85	0.1	4.8	0.2	4.8	0
listener_16	4.4	0.4	2.3	0.2	3.4	1.4	1.2	0	4.6	-0.4	4.3	-0.4	4.1	0
listener_17	3.3	0.6	3.1	0.2	2.6	0.4	3.1	0.8	4.95	0.1	3.7	1.4	3.4	0.8
listener_18	3.35	-0.7	3.15	0.3	1.85	-0.3	2.45	-0.5	4.25	-1.5	2.35	-0.3	2.95	-0.5
listener_19	4.4	1	2.8	-0.4	2.8	1.4	2.6	0.4	4.7	-0.4	4	-1	4.7	-0.2
listener_20	3.6	1	3.4	-1	2.75	0.3	2.65	0.7	4.9	0	3.15	0.7	4.5	0
listener_21	3.45	-1.3	2.3	0.4	1.85	-0.5	1.5	0.2	3	-0.4	2	0.4	1.85	0.1
listener_22	2.5	0.8	3.35	0.3	2.1	1.4	1.95	0.1	4.75	0.1	2.95	1.1	2.15	0.9
listener_23	3.4	-0.8	3.2	1.6	2.3	0.8	2.4	0	4.4	1	3.5	1.8	2.9	-0.6
listener_24	4.05	0.9	4.1	0.8	3.35	-0.1	3.45	0.1	4.95	-0.1	3.95	-0.1	3.6	0.2

Average and difference in scoring for each listener and each trial for B test.

13.4 ANNEX 4: Codec verification

As specified in the test plan, the codecs under test were checked, i.e. bit rate and decoded sequences, by independent parties. This chapter summarizes the results of the verification as reported by the check sites.

13.4.1 Narrowband CELP and Wideband CELP

NTT Human Interface Labs has performed the bitstream verification for the Narrowband and Wideband CELP items. NTT confirmed that all Narrowband CELP and Wideband CELP bitstreams could be decoded successfully. Furthermore, it is confirmed that the bitrate of the CELP18 bitstreams is 18.2 kbps and for the CELP6 streams is 6.0 kbps for all items.

An example of bitrate calculations are shown below.

CELP18

Duration = (wavdata_size - header_size) * 8 /(sampling_freq*16)
= (639446 - 86) *8 /(16000*16)
= 19.98 [sec]
Total Bits =(bitstream_size- header_size)*8
= 45455*8
= 363640 [bits]
Bit rate = Total_Bits/Duration
= 18200.2 [bit/s]

CELP6

Duration = (wavdata_size - header_size) * 8 /(sampling_freq*16)
= (320044 - 44) *8 /(8000*16)
= 20.00 [sec]
Total Bits = (bitstream_size- header_size)*8
= (15226 -226)*8
= 120000 [bits]
Bit rate = Total_Bits/Duration
= 6000.0 [bit/s]

13.4.2 ITU G.723

CSELT performed the verification for the ITU G.723

Decoding and upsampling verification
All the 51 bitstreams have been decoded successfully by the executable decoder included in the MPEG package (for the 4 items available in the package also as reconstructed signals, the corresponding output files are identical). All the 51 reconstructed signals are successfully

upsampled by the executable tool included in the MPEG package (for the 4 items available in the package also as upsampled signals, the corresponding output files are identical).

Bitstream format verification
All the 51 bitstreams are successfully decoded by the official ITU codec (floating point version) available at the ITU web site as source code (the output files are identical to the corresponding ones produced before using the decoder included in the MPEG package).

Bitrate verification
For all the 51 couples of bitstreams and reconstructed signals, the resulting bitrate is exactly 6400 bit/s. This result is compliant to recommendation ITU-T G.723.1, even though the nominal rate is 6300 bit/s. In fact, for each frame (frame length 30 ms), the bitstream uses 192 bits (that is 24 bytes). 189 bits carry the compressed audio information (this figure leads to the nominal rate) and the 3 extra bits are so allocated: 1 bit for the rate selection flag (as the standard is dual-rate, switchable at any frame boundary); 1 bit for the VAD flag; 1 bit unused.

Delay compensation is not taken into account to compute this result. Original audio signal and reconstructed signal result to be delayed of 7.5 ms (which is the codec look-ahead). Therefore the first 7.5 ms of the output signal could be excluded, but this figure is in any case negligible when compared to the whole item duration of about 18 s (value averaged over the 51 items).

13.4.3 Twin-VQ

CCETT had the task to verify the bit streams and decoded sequences for the TwinVQ core. In order to do that zip files containing all the necessary information has been provided by FHG. These archives contain:

- 51 bit streams (result of encoding with codec B5)
- tools for decoding and up-sampling
- 4 decoded sequences at 24 kHz : item16, item26, item36 and item46
- 4 up-sampled sequences at 48 kHz : item16, item26, item36 and item46

The outcomes of the verification are as follows:

a) successful decoding of the 51 bit streams with the decoder that has been provided
b) successful comparison of the decoded sequences for item16, item26, item36 and item46 at 24 kHz.
c) successful comparison of the up-sampled sequences for item16, item26, item36 and item46 at 48 kHz.
d) the bit stream verification gives a total bit rate that goes from 25.018 kbps for item27 to 25.635 kbps for item 41.

Please note that for the bitstream evaluation d) this computation takes the flexmux overhead into account. The overhead is 61 bytes for the header and then 6 bytes per frame (3 bytes AAC 3 bytes TVQ). Philips has conducted a net bit rate verification that gives:

Average average bitrate per item: 24.10 kbps
Minimum average bitrate per item: 23.89 kbps (item 25)
Maximum average bitrate per item: 24.51 kbps (item 41)

The actual bit rate used by the TwinVQ core couldn't be checked with the provided tools at CCETT but Philips has verified that this bit rate is 6 kbps (fixed).

13.4.4 Layer III and Perfect AM

Samsung performed the bitrate verification on layer-3 and perfect AM.

The layer-3 verification obtained the following results:

- All the 51 encoded bitstreams were decoded and compared with provided materials. No differences were found.
- 5 bitstreams (#5, #17, #29, #35 and #47) are upsampled and compared with up-sampled reference. No differences were found. The upsampling program provided by FhG was used.
- The average bitrate was calculated. The average bitrate for the layer-3 bitstreams is 24034.65 kbps. This average bitrate was calculated from 51 items. The bitrate range is between 24028.67 and 24121.83 kbps.

The Perfect-AM frequency analysis obtained the following the results:

- Comparison of the up-sampled results with the provided reference items #5, #17, #29, #35 and #47 did not show any differences.
- The up-sampled results were analyzed by frequency plot using a 4096-point FFT. The maximum values of each frequency were calculated from frames and checked whether the maximum value of each frequency meet the criteria on perfect AM. The results showed that those items met the reference conditions of AM perfect: -3dB at 73Hz and 2400Hz and -50dB at 24Hz and 5300Hz.

13.4.5 AAC at 18 kbps, AAC at 24 kbps, AAC + NB-CELP and AAC + TwinVQ

Philips performed the verification of the AAC_18, AAC_24, AAC_TWINVQ and AAC_CELP bitstreams.

Bitrate verification
Codec B2 (AAC pure, 16 kHz, 18 kbps):
Average average bitrate per item: 18.11 kbps
Minimum average bitrate per item: 17.90 kbps (item25)
Maximum average bitrate per item: 18.36 kbps (item41)
Remark: The decoded signal is delayed for 16 samples.

Codec B3 (AAC pure, 24 kHz, 24 kbps):
Average average bitrate per item: 24.10 kbps
Minimum average bitrate per item: 23.90 kbps (item25)
Maximum average bitrate per item: 24.46 kbps (item41)
Remark: The decoded signal is delayed for 15 samples.

Codec B4 (AAC scal. w. CELP core, 24 kHz, 6 + 18 kbps):
Bitrate per item: 24.00 kbps (fixed bitrate)
Remark: The decoded signal is delayed for 8042 samples.

Codec B5 (AAC scal. w. TwinVQ core, 24 kHz, 6 + 18 kbps):
Average average bitrate per item: 24.10 kbps
Minimum average bitrate per item: 23.89 kbps (item25)
Maximum average bitrate per item: 24.51 kbps (item41)
Remark: The decoded signal is delayed for 4 samples.

General remark
The deviation of the desired bitrate can in all cases be explained by the use of the bit reservoir. For items with a short duration, like item41 (4.67 sec.), this creates the largest deviation from the desired bitrate.

Decoding verification
For each the four codecs each of the 51 bitstreams have been decoded by the executables available on the NADIB CD-ROM. The decoded files are identical to the decoded files available on the CD-ROM.

Upsampling verification
For each four codecs each of the 5 upsampled signals available on the CD-ROM could be reproduced identically by the upsampling program available on the CD-ROM.

Bitstream verification
- The bitstreams of Codec B2 (AAC pure, 16 kHz, 18 kbps) and Codec B3 (AAC pure, 24 kHz, 24 kbps) have been successfully decoded by the latest version of the AAC reference software decoder (980217) and the MPEG4 VM software (V5.0 vm_980211). No problems have been reported.
- The bitstreams of Codec B4 (AAC scaleable with CELP core, 24 kHz, 6 + 18 kbps) have been successfully decoded by the MPEG-4 VM software (V5.0 vm_980211). The core bitstreams (CELP, 8 kHz, 6 kbps) have been extracted and decoded separately. No problems have been reported.
- For decoding the bitstreams of Codec B5 (AAC scaleable with TwinVQ core, 24 kHz, 6 + 18 kbps) no second decoder was available.

13.5 ANNEX 5 : Preselected items for the NADIB test

Comment: You need to set the headings feature so that the second page also has column headings
And, can we identify the gender of the voices in items 28 to 32. At least the question marks should be removed.

#	Filename	duration (sec)	sampler.	Language	Speaker	with	Submitter
1	track05	20	44.1	German	male	-------	DW
2	track06	20	44.1	English	male	-------	DW
3	track07	20	44.1	English	male	-------	DW
4	track08	20	44.1	English	male	-------	DW
5	track11	20	44.1	French	male	-------	DW
6	track17	20	44.1	Chinese	male	-------	DW
7	track21	20	44.1	Japanese	male	-------	DW
8	track25	20	44.1	Arabic	male	-------	DW
9	track26	20	44.1	German	female	-------	DW
10	track31	20	44.1	English	female	-------	DW
11	track32	20	44.1	English	female	-------	DW
12	track33	20	44.1	English	female	-------	DW
13	track35	20	44.1	French	female	-------	DW
14	track40	20	44.1	Chinese	female	-------	DW
15	track43	19	44.1	Japanese	female	-------	DW
16	track47	20	44.1	Arabic	female	-------	DW
17	track53	20	44.1	German	male/female	background music	DW
18	track54	20	44.1	English	male/female	background music	DW
19	track55	20	44.1	English	male/female	background music	DW
20	track57	20	44.1	English	male/female	background music	DW
21	track75	20	44.1	-------	--------	pop	DW
22	track79	20	44.1	-------	--------	folklore	DW
23	track82	20	44.1	-------	--------	classic	DW

24	track86	10	44.1	-------	--------	identification DW	DW
25	track87	10	44.1	-------	--------	news jingle	DW
26	track88	19	44.1	-------	--------	ident. Funkjournal	DW
27	track90	9.97	44.1	-------	--------	ident. WISO	DW
28	hexagon	20	48	French		background music	CCETT
29	jazz	20	48	French		speech followed by jazz	CCETT
30	radiofr1	20	48	French		Radio France mixed speech/music	CCETT
31	radiofr2	17	48	French		-----	CCETT
32	rfi1	20	48	French		Radio France international: news,jingels, mixed	CCETT
33	app_guit	19	24			complex sound +applause	Berkom
34	mussorg	20	24			complex sound +applause	Berkom
35	rock	20	24			complex sound	Berkom
36	suz_24_mono	20	24			female English singing	Sony
37	Sade3_mono	20	24			Pop (English)	Sony
38	Vivaldi_24	20	24			classic	Sony
39	jb01	9.66	48	Japanese	male	background music	Sony
40	jb02	6.65	48	Japanese	female	background noise	Sony
41	jm01	4.67	48	Japanese	male/female	multiple speakers	Sony
42	JazzSwe	20	48			jazz	SR
43	SPEngMale	20	48	English	male		SR
44	SPEngMaEF	20	48	English	male	background sound (sport spectators)	SR
45	SPFraMale	20	48	French	male		SR
46	SPFraMaEF	20	48	French	male	background sound (sport spectators)	SR
47	SPGerMale	20	48	German	male		SR
48	SPGerMaEF	20	48	German	male	background sound (sport spectators)	SR
49	SPSweMale	20	48	Swedish	male		SR
50	SPSweMaEF	20	48	Swedish	male	background sound (sport spectators)	SR
51	SPSweFem	20	48	Swedish	female		SR

13.6 ANNEX 6 : Selection Panel report

Report of the Selection Panel of the NADIB test

The preselection ran without any complication. We in the group had the same opinion about how to listen and how to choose the critical items. Only some details differed but after a discussion we agreed.

We had a very good help from our own software which ran without problems. We had all 51 items available and always 3 items on the screen. It was easy to pick up the different items and codecs to check the sound and artefacts. We had always the original reference available in case we wanted to compare it with the codecs. When we listened to the codecs there was a loop that started every codec after each other for every item. Then we had to start the new item manually. We used two listening conditions: one with loudspeakers and the other with headphones. The listening rooms were quite silent. If one or two of us wanted to detect more details we could go into the other room.

Equipment used for the loudspeaker-room:

Loudspeaker: Genelec 1024
Analogue mixer SATT SAM 82
Computer Dell Pentium 11 333 MHz
Soundcard ADB Multiwav Digital
D/A converter Prism DA1

Headphone-room

Headphones Stax Lambda Pro Driver SRM-Monitor
D/A converter Philips DAC 960
Computer Pentium 11 266 MHz
Soundcard Zefiro ZA2

The computers were acoustically isolated from the listening room.

Some remarks:

We wanted to follow the recommendation how to select different kind of music and speech. It was no problems for us to do that.
For the Swedish (Teracom-test) we chose an English speaking female instead of the Swedish Female speech. There were almost no artefacts on the Swedish (Comment: is this correct?)female speech.

For the French (CCETT) it was important that the French male/female were speaking native French and no dialect.
For the folklore there was only one item and it was critical

Here are the results:

We recommend these items for the A and B-TEST
We also recommend 2 orders for the 10 items.

First Order
1 No 20
2 No 2
3 No 38
4 Native Item Swedish No 36 French 13
5 No 35
6 Native Item Swedish No 50 French 28
7 No 21
8 No 10
9 No 44
10 No 22

Second Order
1 No 38
2 No 10
3 No 21
4 No 44
5 No 20
6 No 22
7 Native Item Swedish No 50 French 28
8 No 2
9 No 35
10 Native Item Swedish No 36 French 13

TRAINING ITEMS

For the A-Test we recommend :
1 No 3
2 No 12
3 No 39
4 No 48

For the B-Test we recommend
1 No 3
2 No 12
3 No 39
4 No 48
5 No 23
6 No 31
7 No 33

13.7 ANNEX 7 : Instructions for scoring and vote sheets

13.7.1 Official English version

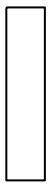
How to perform the listening test

1. Familiarisation or Training phase

The purpose of the training phase is to allow you, as a listener, to identify and become familiar with distortions and artefacts produced by the systems under test. The sound excerpts in the training phase are selected to illustrate the whole range of qualities that may be heard. This fact does NOT necessarily mean that you should give grade 1.0 to the sound excerpt with lowest quality, nor grade 5.0 to the sound excerpt with highest quality. You should use the range you find appropriate. During the training phase you will also become familiar with the test procedure. After the training, you should know what to listen for and how to grade the quality of the excerpts, and will then proceed with the real test.

During the training phase, you will hear both the reference (original), A, and the processed versions, B, of each item of audio material, presented in the sequence A-B-A-B. Announcements on the screen will remind you whether you are going to listen to the reference (A) or to the processed version (B). The duration of the audio sequences will typically be between 15 and 25 seconds.

You should use the quality scale as follows



5.0 Excellent
4.0 Good
3.0 Fair
2.0 Poor
1.0 Bad

You are advised to use the reference (A) stimulus as an indication of the optimum quality for each programme item, i.e. it corresponds to ”Excellent”. The grading scale is continuous from 5.0 to 1.0, and you should give your answer to an accuracy of one decimal place e.g. 3.2, 1.9.

Whilst you should be considering during the training phase how you, as an individual, will interpret the audible impairments in terms of the grading scale, it is important that you should not discuss this personal interpretation with the other subjects at any time.

All grades given during the training phase will be disregarded.

2. Grading phase

The purpose of the test is to grade the quality of the audio material you will hear.

For each item, you will listen to two versions of a given audio excerpt. The versions will be identified as A - the reference and B - the processed version, and will be presented in the sequence A-B-A-B. Afterwards there will be 8 seconds of silence during which you write down

your judgement of the quality level of B. If you like, you can write down a comment as well, indicating, perhaps, why you gave the grade you did. After this silent period the next item starts with an aural announcement indicating the number of the new item: ”item nn”. Each session will contain approximately 15 items to be graded.

Test site :
Session N° :
Random N° :
Date :
Name :
Age :
Profession :
Expert / Non expert :

5.0	Excellent
4.0	Good
3.0	Fair
2.0	Poor
1.0	Bad

The quality scale

You should grade your evaluations to an accuracy of one decimal place.

# item	Grade of B	Comments
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		

13.7.2 TERACOM version

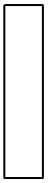
How to perform the listening test

1. Familiarisation or Training phase

The purpose of the training phase is to allow you, as a listener, to identify and become familiar with distortions and artefacts produced by the systems under test. The sound excerpts in the training phase are selected to illustrate the whole range of qualities that may be heard during the real test This fact does NOT necessarily mean that you should give grade 1.0 to the sound excerpt with lowest quality, nor grade 5.0 to the sound excerpt with highest quality. You should use the range you find appropriate. During the training phase you will also become familiar with the test procedure. After the training, you should know what to listen for and how to grade the quality of the excerpts, and will then proceed with the real test.

During the training phase, you will hear both the reference (original), A, and the processed versions, B, of each item of audio material, presented in the sequence A-B-A-B. Announcements on the screen will remind you whether you are listening to the reference (A) or to the processed version (B). The duration of the audio sequences will typically be between 15 and 25 seconds.

You should use the quality scale as follows



5.0 Mycket bra
4.0 Bra
3.0 Varken bra eller dålig
2.0 Dålig
1.0 Mycket dålig

You are advised to use the reference (A) stimulus as an indication of the optimum quality for each programme item, i.e. it corresponds to ”Excellent”. The grading scale is continuous from 5.0 to 1.0, and you should give your answer to an accuracy of one decimal place e.g. 3.2 or 1.9.

Whilst you should be considering during the training phase how you, as an individual, will interpret the audible impairments in terms of the grading scale, it is important that you should not discuss this personal interpretation with the other listeners at any time.

All grades given during the training phase will be disregarded.

2. Grading phase

The purpose of the test is to grade the quality of the audio material you will hear.

For each item, you will listen to two versions of a given audio excerpt. The versions will be identified as A - the reference and B - the processed version, and will be presented in the sequence A-B-A-B. Afterwards there will be 8 seconds of silence during which you write down your judgement of the quality level of B. If you like, you can write down a comment as well, indicating, perhaps, the reasons for a specific grade. After this silent period the next item starts with an aural announcement indicating the number of the new item: ”item nn”. Each session will contain approximately 15 items to be graded.

Test site :

Session No. :

Random No :

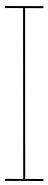
Date :

Name :

Age:

Profession

Expert/Non expert:



5.0 Mycket bra
4.0 Bra
3.0 Varken bra eller dålig
2.0 Dålig
1.0 Mycket Dålig

The quality scale

You should grade your evaluations to an accuracy of one decimal place.

# item	Grade of B	Comments
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		

13.7.3 CCETT Version

Comment effectuer le test d'écoute

1. La phase de familiarisation ou d'entraînement

L'objectif de la phase d'entraînement est de vous permettre en tant qu'auditeur de vous familiariser avec les distorsions et les défauts produits par les systèmes que l'on évalue. Les extraits sonores que vous écoutez pendant la phase d'entraînement sont choisis pour illustrer toute la gamme de qualité que l'on pourrait entendre. Cela ne signifie pas que vous devez donner une note de 1.0 à l'extrait qui a la qualité la plus basse ni 5.0 à celui qui à la qualité la plus élevée. Il faut que vous utilisiez la gamme qui vous semble appropriée. Pendant la phase d'entraînement, vous vous familiariserez aussi avec la procédure de test. Après la phase d'entraînement, vous devriez être en mesure d'apprécier l'échelle de qualité des séquences, et ensuite d'être préparé pour le test réel.

Au cours de la phase d'entraînement, vous écouterez la référence (original), A, et les versions traitées, B, de chaque extrait sonore, présenté selon le séquencement A-B-A-B. Des annonces visuelles sur écran vous rappelleront si vous écoutez soit la référence (A) soit la version traitée (B). La durée des séquences sonores sera typiquement d'entre 15 et 25 secondes.

Il faudra utiliser l'échelle de qualité suivante :

5.0	Excellent
4.0	Bon
3.0	Assez bon
2.0	Médiocre
1.0	Mauvais

Il est conseillé d'utiliser le stimulus de référence (A) comme une indication de la qualité optimale pour chaque programme, c'est-à-dire qu'il correspond à « Excellent ». L'échelle de notation est continue de 5.0 à 1.0, et vous devriez donner votre note à une décimale près, par exemple 3.2, 1.9.
Pendant la phase d'entraînement, il faut que vous pensiez individuellement à la façon dont vous interpréterez les dégradations audibles en fonction de l'échelle de notation. Il est important que vous ne discutiez pas de cette interprétation personnelle avec les autres auditeurs.
Les notes données pendant la phase d'entraînement ne seront pas prises en compte.

2. La phase de notation

L'objectif du test est de noter la qualité des matériaux sonores que vous écouterez.
Vous entendrez deux versions d'un extrait sonore pour chaque séquencement. Les versions seront identifiées comme A - la référence, et B - la version traitée, et seront présentées selon le séquencement A - B - A - B. Ensuite il y aura un silence de 8 secondes pendant lequel vous écrirez votre évaluation de la qualité de B. Si vous le voulez, vous pouvez écrire une remarque donnant peut-être une raison pour laquelle vous avez donné cette note. Après cette période de silence la séquence suivante commencera avec une annonce orale indiquant le numéro de la nouvelle séquence: «séquence N°». Chaque session contiendra approximativement 15 extraits sonores à évaluer.

Session N°
Aléatoire N°
Date
Nom
Age
Profession
Expert / Amateur

4.0 Bon
3.0 Assez bon
2.0 Médiocre
1.0 Mauvais

L'échelle de qualité

Vous devez donner votre évaluation à une décimale près :

Extrait N°	Note de B	Commentaires
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		

13.8 ANNEX 8 : List of the "pseudo-randomisation" of each test

A Test

Random 1	**Random 2**	**Random 3**	**Random 4**
Session 1	Session 1	Session 1	Session 1
codeca2/item20	codeca2/item38	codeca1/item44e	codeca3/item13
codeca1/item28	codeca3/item10	codeca3/item2	codeca3/item35
codeca2/item38	codeca1/item21	codeca2/item10	codeca2/item2
codeca3/item13	codeca1/item13	codeca1/item28	codeca1/item20
codeca2/item35	codeca3/item20	codeca3/item13	codeca2/item28
codeca3/item28	codeca2/item44e	codeca2/item35	codeca2/item22
codeca1/item2	codeca3/item21	codeca1/item38	codeca3/item20
codeca3/item10	codeca3/item38	codeca1/item13	codeca3/item2
codeca1/item44e	codeca2/item10	codeca2/item2	codeca1/item44e
codeca3/item2	codeca1/item44e	codeca2/item20	codeca3/item21
codeca3/item20	codeca2/item2	codeca3/item22	codeca2/item10
codeca2/item13	codeca1/item10	codeca2/item44e	codeca1/item38
codeca3/item22	codeca2/item28	codeca1/item35	codeca2/item13
codeca1/item38	codeca3/item2	codeca3/item21	codeca2/item35
codeca2/item10	codeca1/item20	codeca3/item28	**codeca3/item2**
codeca3/item35	codeca2/item13	codeca1/item10	codeca1/item28
codeca2/item28	codeca1/item2	codeca2/item13	codeca1/item22
Session 2	Session 2	Session 2	Session 2
codeca2/item21	codeca3/item28	codeca3/item38	codeca2/item44e
codeca1/item44e	codeca1/item22	**codeca1/item44e**	codeca2/item21
codeca2/item13	codeca3/item35	**codeca3/item2**	**codeca1/item44e**
codeca1/item21	**codeca2/item13**	codeca3/item20	codeca3/item10
codeca2/item22	codeca1/item38	codeca2/item22	codeca2/item38
codeca1/item20	**codeca1/item44e**	codeca1/item21	codeca1/item13
codeca2/item44e	codeca2/item35	codeca3/item44e	codeca1/item35
codeca3/item38	codeca3/item44e	codeca2/item38	codeca1/item2
codeca3/item21	codeca2/item20	codeca3/item10	**codeca2/item13**
codeca1/item35	codeca3/item22	codeca1/item22	codeca3/item28
codeca2/item2	codeca1/item28	codeca2/item21	codeca3/item22
codeca1/item13	**codeca3/item2**	**codeca2/item13**	codeca2/item20
codeca1/item10	codeca2/item22	codeca3/item35	codeca3/item44e
codeca3/item44e	codeca1/item35	codeca2/item28	codeca1/item21
codeca3/item2	codeca2/item21	codeca1/item2	codeca1/item10
codeca1/item22	codeca3/item13	codeca1/item20	codeca3/item38

Test B

Random 1	**Random 2**	**Random 3**	**Random 4**
Session 1	Session 1	Session 1	Session 1
codecb6/item20	codecb3/item38	codecb7/item22	codecb4/item13
codecb1/item2	codecb7/item10	codecb4/item44e	codecb5/item35
codecb3/item38	codecb4/item21	codecb6/item10	codecb6/item2
codecb4/item13	codecb5/item44e	codecb7/item21	codecb7/item28
codecb6/item35	codecb1/item20	codecb5/item28	codecb1/item22
codecb2/item28	codecb7/item22	codecb3/item35	codecb4/item20
codecb6/item21	codecb2/item28	codecb2/item13	codecb3/item44e
codecb5/item10	codecb3/item2	codecb7/item38	codecb7/item21
codecb7/item44e	codecb1/item35	codecb6/item2	codecb3/item10
codecb3/item22	codecb4/item13	codecb2/item20	codecb1/item38
codecb2/item20	codecb5/item38	codecb4/item22	codecb2/item13
codecb7/item2	codecb6/item10	codecb5/item44e	codecb3/item35
codecb5/item38	codecb2/item21	codecb1/item10	codecb1/item2
codecb2/item13	codecb7/item44e	codecb2/item21	codecb6/item28
codecb4/item35	codecb2/item20	codecb4/item28	codecb7/item22
codecb4/item28	codecb2/item22	codecb7/item35	codecb2/item20
Session 2	Session 2	Session 2	Session 2
codecb3/item21	codecb4/item28	codecb3/item13	codecb1/item44e
codecb2/item10	codecb1/item2	codecb6/item38	codecb5/item21
codecb1/item44e	codecb3/item35	codecb5/item2	codecb6/item10
codecb5/item22	codecb2/item13	**codecb2/item20**	**codecb1/item38**
codecb7/item20	codecb6/item38	codecb6/item22	codecb5/item13
codecb6/item2	codecb5/item10	codecb3/item44e	codecb2/item35
codecb1/item38	codecb7/item21	codecb2/item10	codecb4/item2
codecb3/item13	codecb2/item44e	codecb4/item21	codecb3/item28
codecb7/item35	codecb6/item20	codecb1/item28	**codecb7/item22**
codecb3/item28	codecb4/item22	codecb5/item35	codecb3/item20
codecb6/item21	codecb3/item28	codecb6/item13	codecb6/item44e
codecb4/item10	codecb7/item2	codecb1/item38	codecb1/item21
codecb6/item44e	codecb5/item35	codecb3/item2	codecb2/item10
codecb1/item22	codecb1/item13	codecb5/item20	codecb7/item38
codecb2/item20	codecb4/item38	**codecb7/item22**	codecb1/item13
Session 3	Session 3	Session 3	Session 3
codecb2/item2	**codecb5/item10**	codecb1/item44e	codecb4/item35
codecb4/item38	codecb3/item21	codecb3/item10	codecb3/item2

Test B

Random 1	**Random 2**	**Random 3**	**Random 4**

codecb7/item13	codecb6/item44e	codecb6/item21	codecb4/item28
codecb5/item35	codecb4/item20	codecb2/item28	codecb5/item22
codecb7/item28	**codecb7/item22**	codecb6/item35	codecb1/item20
codecb2/item21	codecb6/item28	codecb7/item13	codecb7/item44e
codecb6/item10	codecb5/item2	codecb5/item38	codecb3/item21
codecb3/item44e	codecb4/item35	codecb4/item2	codecb5/item10
codecb6/item22	codecb5/item13	codecb1/item20	codecb6/item38
codecb5/item20	codecb7/item38	codecb5/item22	codecb7/item13
codecb4/item2	codecb1/item10	codecb7/item44e	**codecb3/item35**
codecb1/item38	codecb6/item21	codecb4/item10	codecb7/item2
codecb6/item13	codecb3/item44e	codecb3/item21	codecb1/item28
codecb3/item35	codecb7/item20	**codecb4/item28**	codecb2/item22
codecb1/item28	codecb5/item22	codecb2/item35	codecb7/item20
Session 4	Session 4	Session 4	Session 4
codecb7/item21	codecb7/item28	codecb1/item13	codecb5/item44e
codecb3/item10	codecb6/item2	codecb3/item38	codecb6/item21
codecb4/item44e	**codecb3/item35**	codecb7/item2	codecb1/item10
codecb7/item22	codecb6/item13	codecb4/item20	codecb4/item38
codecb3/item20	codecb1/item38	codecb1/item22	codecb3/item13
codecb5/item2	codecb2/item10	codecb2/item44e	codecb7/item35
codecb2/item38	codecb5/item21	codecb5/item10	codecb5/item2
codecb1/item13	codecb1/item44e	**codecb6/item21**	codecb2/item28
codecb3/item35	codecb5/item20	codecb7/item28	codecb6/item22
codecb6/item28	codecb3/item22	**codecb3/item35**	**codecb2/item20**
codecb4/item21	**codecb4/item28**	codecb4/item13	codecb4/item44e
codecb5/item10	codecb2/item2	codecb2/item38	codecb2/item21
codecb7/item22	codecb6/item35	codecb1/item2	codecb4/item10
codecb4/item20	codecb3/item13	codecb7/item20	codecb5/item38
codecb5/item44e	codecb2/item38	codecb3/item22	codecb6/item13
Session 5	Session 5	Session 5	Session 5
codecb3/item2	codecb4/item10	codecb6/item44e	codecb6/item35
codecb7/item38	**codecb6/item21**	**codecb5/item10**	codecb2/item2
codecb5/item13	codecb4/item44e	codecb1/item21	**codecb4/item28**
codecb1/item35	codecb3/item20	codecb6/item28	codecb3/item22
codecb4/item28	codecb1/item22	codecb1/item35	codecb5/item20

Test B

Random 1	**Random 2**	**Random 3**	**Random 4**

codecb5/item21	codecb5/item28	codecb5/item13	codecb2/item44e
codecb1/item10	codecb4/item2	codecb4/item38	**codecb6/item21**
codecb2/item44e	codecb2/item35	codecb2/item2	codecb7/item10
codecb4/item22	codecb7/item13	codecb3/item20	codecb3/item38
codecb1/item20	**codecb1/item38**	codecb2/item22	codecb1/item35
codecb6/item38	codecb3/item10	codecb7/item10	codecb5/item28
codecb2/item35	codecb1/item21	codecb5/item21	codecb4/item22
codecb5/item28	**codecb2/item20**	codecb3/item28	codecb6/item20
codecb1/item21	codecb6/item22	codecb4/item35	codecb4/item21
codecb7/item10	codecb1/item28	**codecb1/item38**	**codecb5/item10**
codecb2/item22	codecb7/item35	codecb6/item20	codecb2/item38

13.9 ANNEX 9 : Tables resulting from the test result analysis

Table 1: Reliability of each subject, showing mean and CI of difference between scores in repeated trials. Highlighted subjects had CIs extending outside [-1,1] and were eliminated.

	Mean	*Lower CI*	*Upper CI*
14			
C1	-2.0000E-02	-.4919	.4519
C10	.3200	-.1579	.7979
C11	-.2700	-.6270	8.696E-02
C12	.1200	-.4764	.7164
C13	.1200	-.2715	.5115
C14	-2.0000E-02	-.5389	.4989
C15	-.1900	-.5195	.1395
C16	-.1600	-.5266	.2066
C17	-.4400	-.7776	-.1024
C18	.3300	-9.3949E-02	.7539
C19	3.000E-02	-.5393	.5993
C2	4.000E-02	-.3172	.3972
C20	-.3000	-.7144	.1144
C21	.1000	-.3061	.5061
C22	-.3700	-.8398	9.976E-02
C23	-.4600	-1.0973	.1773
C24	3.000E-02	-.3966	.4566
C3	-.3100	-.8635	.2435
C4	-3.0000E-02	-.5381	.4781
C5	-.1400	-.3789	9.893E-02
C6	8.000E-02	-.2922	.4522
C9	8.000E-02	-.2717	.4317
T1	-.6400	-1.0533	-.2267
T10	-5.0000E-02	-.3910	.2910
T11	.2200	-5.1459E-02	.4915
T12	2.000E-02	-.5119	.5519
T13	-.1200	-.3786	.1386
T14	-.6500	-1.2105	-8.9510E-02
T15	-.1200	-.5516	.3116
T16	-4.0000E-02	-.3265	.2465
T17	-7.0000E-02	-.4560	.3160
T18	-.1000	-.4439	.2439
T19	2.000E-02	-.4777	.5177
T2	.1500	-.2289	.5289
T20	6.000E-02	-.3927	.5127
T21	-.2600	-.7051	.1851
T22	-9.0000E-02	-.7002	.5202
T23	-.1500	-.5289	.2289
T24	.3400	7.602E-04	.6792
T25	.5000	-.1617	1.1617
T26	-.1600	-.4281	.1081
T27	-.1000	-.4054	.2054
T28	.2300	-7.9161E-02	.5392
T29	-.1700	-.5254	.1854
T3	4.000E-02	-.2302	.3102
T30	-5.0000E-02	-.3366	.2366
T31	-.1900	-.3821	2.098E-03
T32	-.3900	-.7297	-5.0341E-02
T33	2.220E-17	-.3521	.3521
T34	-.1500	-.6550	.3550
T35	-.3700	-.6988	-4.1230E-02
T36	-3.0000E-02	-.4553	.3953
T4	.1600	-.1016	.4216
T5	.2700	8.680E-03	.5313
T6	.2000	-.2435	.6435
T7	-.1600	-.6431	.3231
T8	.1100	-.1690	.3890
T9	7.000E-02	-.5609	.7009

Table 2: Codec-by-codec results

SITE	*CODEC*	*Mean*	*CI Lower*	*CI Upper*
C	Twin-VQ 6 kbps	1.7676	1.6699	1.8653
	NB-CELP 6 kbps	2.5590	2.4371	2.6810
	G.723.1 6.3 kbps	2.7105	2.5817	2.8392
	WB-CELP 18.2 kbps	2.2414	2.1170	2.3658
	Perfect AM	2.7971	2.6790	2.9153
	AAC 18 kbps	3.0790	2.9592	3.1988
	AAC scal w/ TwinVQ 24 kbps	3.4371	3.3316	3.5426
	MPEG-2 Layer III 24 kbps	3.5814	3.4692	3.6936
	AAC scal w/CELP 24 kbps	3.6724	3.5574	3.7874
	AAC 24 kbps	4.1305	4.0198	4.2412
T	Twin-VQ 6 kbps	2.0173	1.9324	2.1021
	NB-CELP 6 kbps	2.6679	2.5661	2.7697
	G.723.1 6.3 kbps	2.8612	2.7570	2.9654
	WB-CELP 18.2 kbps	2.3712	2.2743	2.4681
	Perfect AM	2.8333	2.7520	2.9147
	AAC 18 kbps	3.2879	3.1937	3.3821
	MPEG-2 Layer III 24 kbps	3.4994	3.4186	3.5802
	AAC scal w/ TwinVQ 24 kbps	3.5536	3.4754	3.6319
	AAC scal w/CELP 24 kbps	3.7273	3.6376	3.8170
	AAC 24 kbps	4.1367	4.0504	4.2229

Table 3: Item-by-item performance of each codec, pooled for the two test sites.

ITEM	*CODEC*	*Mean*	*CI Lower*	*CI Upper*
2	Twin-VQ 6 kbps	1.5889	1.4488	1.7289
	G.723.1 6.3 kbps	3.4500	3.2715	3.6285
	NB-CELP 6 kbps	3.3759	3.2008	3.5511
	AAC scal w/CELP 24 kbps	3.1537	2.9566	3.3508
	MPEG-2 Layer III 24 kbps	3.1500	2.9737	3.3263
	Perfect AM	2.9500	2.7486	3.1514
	AAC 24 kbps	3.5889	3.3683	3.8095
	WB-CELP 18.2 kbps	3.0796	2.8593	3.3000
	AAC scal w/ TwinVQ 24 kbps	3.1019	2.9266	3.2771
	AAC 18 kbps	2.5315	2.3727	2.6903
10	Twin-VQ 6 kbps	1.7759	1.6247	1.9272
	G.723.1 6.3 kbps	3.3407	3.1694	3.5121
	NB-CELP 6 kbps	3.2037	3.0094	3.3980
	AAC scal w/CELP 24 kbps	3.0926	2.8878	3.2974
	MPEG-2 Layer III 24 kbps	3.1741	2.9704	3.3777
	Perfect AM	3.1889	2.9831	3.3947
	AAC 24 kbps	3.4593	3.2391	3.6794
	WB-CELP 18.2 kbps	3.0796	2.8800	3.2792
	AAC scal w/ TwinVQ 24 kbps	3.1796	2.9966	3.3626

	AAC 18 kbps	2.4778	2.2889	2.6666
13	Twin-VQ 6 kbps	1.6095	1.4075	1.8115
	G.723.1 6.3 kbps	3.4667	3.1636	3.7698
	NB-CELP 6 kbps	3.0190	2.7135	3.3246
	AAC scal w/CELP 24 kbps	3.3286	3.0472	3.6100
	MPEG-2 Layer III 24 kbps	3.3762	2.9836	3.7688
	Perfect AM	3.0810	2.7013	3.4606
	AAC 24 kbps	3.9000	3.6242	4.1758
	WB-CELP 18.2 kbps	2.7429	2.4737	3.0120
	AAC scal w/ TwinVQ 24 kbps	3.1905	2.8848	3.4961
	AAC 18 kbps	2.4000	2.0987	2.7013
20	Twin-VQ 6 kbps	1.6907	1.5544	1.8271
	G.723.1 6.3 kbps	3.1500	2.9504	3.3496
	NB-CELP 6 kbps	2.9907	2.7967	3.1848
	AAC scal w/CELP 24 kbps	3.8463	3.6503	4.0423
	MPEG-2 Layer III 24 kbps	3.4796	3.2858	3.6735
	Perfect AM	2.6630	2.4474	2.8785
	AAC 24 kbps	4.2019	4.0195	4.3842
	WB-CELP 18.2 kbps	2.4093	2.2265	2.5920
	AAC scal w/ TwinVQ 24 kbps	3.4000	3.2058	3.5942
	AAC 18 kbps	2.9444	2.7419	3.1469
21	Twin-VQ 6 kbps	1.9185	1.7326	2.1045
	G.723.1 6.3 kbps	2.5148	2.3105	2.7192
	NB-CELP 6 kbps	2.2463	2.0467	2.4459
	AAC scal w/CELP 24 kbps	3.8759	3.6743	4.0775
	MPEG-2 Layer III 24 kbps	3.7111	3.5103	3.9119
	Perfect AM	2.6204	2.3798	2.8610
	AAC 24 kbps	4.6444	4.5135	4.7754
	WB-CELP 18.2 kbps	1.9574	1.7702	2.1446
	AAC scal w/ TwinVQ 24 kbps	3.5815	3.3964	3.7665
	AAC 18 kbps	3.2759	3.0615	3.4904
22	Twin-VQ 6 kbps	2.2667	2.0582	2.4751
	G.723.1 6.3 kbps	1.9037	1.7031	2.1043
	NB-CELP 6 kbps	1.6870	1.5146	1.8595
	AAC scal w/CELP 24 kbps	4.1222	3.9309	4.3135
	MPEG-2 Layer III 24 kbps	3.7852	3.5833	3.9870
	Perfect AM	2.6389	2.4354	2.8423
	AAC 24 kbps	4.5148	4.3724	4.6572
	WB-CELP 18.2 kbps	1.4870	1.3321	1.6420
	AAC scal w/ TwinVQ 24 kbps	3.8111	3.6301	3.9921
	AAC 18 kbps	3.6870	3.5088	3.8653
28	Twin-VQ 6 kbps	1.7429	1.5096	1.9761
	G.723.1 6.3 kbps	2.9095	2.6056	3.2135
	NB-CELP 6 kbps	2.9524	2.6171	3.2877
	AAC scal w/CELP 24 kbps	3.7333	3.3763	4.0904
	MPEG-2 Layer III 24 kbps	3.6476	3.3265	3.9687
	Perfect AM	2.9952	2.6493	3.3411

	AAC 24 kbps	4.0048	3.6748	4.3347
	WB-CELP 18.2 kbps	2.1381	1.8671	2.4091
	AAC scal w/ TwinVQ 24 kbps	3.3905	3.0686	3.7123
	AAC 18 kbps	2.9048	2.6092	3.2003
35	Twin-VQ 6 kbps	1.9111	1.7198	2.1024
	G.723.1 6.3 kbps	2.5852	2.3278	2.8426
	NB-CELP 6 kbps	2.2019	1.9670	2.4367
	AAC scal w/CELP 24 kbps	4.0722	3.8802	4.2642
	MPEG-2 Layer III 24 kbps	3.8630	3.6443	4.0817
	Perfect AM	2.5926	2.3599	2.8253
	AAC 24 kbps	4.4963	4.3193	4.6733
	WB-CELP 18.2 kbps	2.0296	1.8041	2.2552
	AAC scal w/ TwinVQ 24 kbps	3.8204	3.6107	4.0301
	AAC 18 kbps	3.8667	3.6815	4.0519
36	Twin-VQ 6 kbps	1.8000	1.6134	1.9866
	G.723.1 6.3 kbps	3.2394	2.9871	3.4917
	NB-CELP 6 kbps	2.7545	2.5423	2.9668
	AAC scal w/CELP 24 kbps	4.0455	3.7688	4.3221
	MPEG-2 Layer III 24 kbps	3.7727	3.5003	4.0451
	Perfect AM	2.6970	2.4869	2.9070
	AAC 24 kbps	4.3788	4.1297	4.6279
	WB-CELP 18.2 kbps	2.1455	1.9054	2.3855
	AAC scal w/ TwinVQ 24 kbps	3.6909	3.4415	3.9403
	AAC 18 kbps	3.6848	3.4116	3.9581
38	Twin-VQ 6 kbps	3.1907	2.9661	3.4153
	G.723.1 6.3 kbps	1.5537	1.4001	1.7073
	NB-CELP 6 kbps	1.6389	1.4987	1.7791
	AAC scal w/CELP 24 kbps	4.2907	4.1234	4.4581
	MPEG-2 Layer III 24 kbps	3.9704	3.8072	4.1335
	Perfect AM	2.9407	2.7513	3.1302
	AAC 24 kbps	4.6741	4.5418	4.8063
	WB-CELP 18.2 kbps	1.4222	1.3004	1.5440
	AAC scal w/ TwinVQ 24 kbps	4.0963	3.9297	4.2629
	AAC 18 kbps	4.0352	3.8574	4.2130
44	Twin-VQ 6 kbps	1.4796	1.3551	1.6041
	G.723.1 6.3 kbps	3.2056	2.9926	3.4185
	NB-CELP 6 kbps	3.1037	2.8737	3.3337
	AAC scal w/CELP 24 kbps	3.3500	3.1167	3.5833
	MPEG-2 Layer III 24 kbps	3.1556	2.9423	3.3688
	Perfect AM	2.8259	2.5924	3.0594
	AAC 24 kbps	3.8704	3.6480	4.0927
	WB-CELP 18.2 kbps	2.8185	2.5929	3.0442
	AAC scal w/ TwinVQ 24 kbps	3.3093	3.1199	3.4986
	AAC 18 kbps	3.3241	3.1348	3.5134
50	Twin-VQ 6 kbps	1.5970	1.4284	1.7655
	G.723.1 6.3 kbps	3.0485	2.7162	3.3808

	NB-CELP 6 kbps	2.9485	2.6828	3.2141
	AAC scal w/CELP 24 kbps	3.3333	3.0630	3.6036
	MPEG-2 Layer III 24 kbps	3.2515	3.0356	3.4674
	Perfect AM	2.8818	2.5918	3.1718
	AAC 24 kbps	3.5061	3.2384	3.7738
	WB-CELP 18.2 kbps	2.8061	2.5400	3.0721
	AAC scal w/ TwinVQ 24 kbps	3.2212	2.9798	3.4626
	AAC 18 kbps	2.6333	2.3884	2.8783