INTERNATIONAL ORGANISATION FOR STANDARDISATION

ORGANISATION INTERNATIONALE DE NORMALISATION

ISO/IEC JTC1/SC29/WG11

CODING OF MOVING PICTURES AND ASSOCIATED AUDIO

ISO/IEC JTC1/SC29/WG11 **N1419**

November 1996

Report on the Formal Subjective Listening Tests of MPEG-2 NBC multichannel audio coding.

David Kirby, BBC R&D, Kingswood Warren, UK.
Kaoru Watanabe, NHK, Tokyo, Japan.

Status: Approved

This report has been compiled with the assistance of:

Alan Kimber & Peter Williams The University of Guildford, Surrey, UK.

Preface

This document reports on the preparations for, and presents the results of, the subjective listening tests on the MPEG-2 “NBC” coding algorithm carried out in Kingswood Warren, UK, and Tokyo, Japan, between 16 September and 11 October 1996.

The following additional contributions are also included:

- Selection Panel Report
- Instructions to listeners
- Test conditions and equipment
- Statistical procedures

Acknowledgements

The authors of this report would like to acknowledge gratefully the work and assistance of the following people:

- all the listeners who participated in the tests
- the selection panel for their excellent work

  Thomas Buchholz, Telekom TZD Berlin, Germany

  Kazuho Ono, NHK Science and Technical Research Labs, Tokyo, Japan

  Andrew McParland, BBC Research and Development Department, UK

  John Fletcher, BBC Research and Development Department, UK
- Peter Schreiner (Scientific Atlanta), Chairman of the MPEG Audio Subgroup, for advice and support throughout these tests.
- at the BBC: Ken Taylor and David Meares and at NHK: Satoru Koizumi for their extensive and valuable support during the preparation and test phases.
- Ron Burns (Hughes Electronics), Mike Coleman (Five Bats), Alberto Duenas (DMV), Frank Feige (Deutsche Telekom), Hendrik Fuchs (University of Hanover) and Jens Spille (DTB) for performing the necessary codec verification work.
- Alan Kimber and Peter Williams, (Mathematics and Computing Science Department, The University of Guildford, Surrey, UK) for performing the statistical analysis of the test data.

Contents

1. Introduction

1.1 Background

In March 1994, Deutsche Telekom and the BBC reported the results of formal listening tests on the MPEG-2 Backwards Compatible multichannel coding algorithms [1]. Eight codecs were evaluated at that time: six MPEG-2 Backwards Compatible (BC)[1] implementations and two Non-Backwards Compatible (NBC) codecs. The results indicated that all of the codecs tested were not acceptable for high quality applications at the tested bitrates. It was also observed that the BC codecs did not perform as well as the NBC codecs at the same bitrate.

As a result of those findings, MPEG decided on two courses of action: firstly, to include, in the proposed MPEG-2 audio standard, additional features which would deliver better audio quality and, secondly, to initiate the development of a Non-Backwards Compatible coding technique. The first of these, together with general improvements to the codec implementations, has led to the improved performance of the MPEG-2 BC codecs, reported by a series of subjective tests [2,3,4].

The development of the MPEG-2 NBC coding technique has proceeded over the last two years and during that time has been subject to various stages of optimisation [5 to 11] based on a combination of technical developments and proving core experiments [12 to 15]. In the early stages, the developments were based on monophonic embodiments, whilst more recently both two-channel and multichannel implementations have been produced.

This programme of work has reached the stage where formal testing of the multichannel implementation is appropriate. Accordingly, at the July 1996 meeting of MPEG in Tampere, the BBC and NHK were jointly charged to conduct formal subjective tests aimed at quantifying the performance of MPEG-2 Non-Backwards Compatible audio codecs operating in a multichannel mode [16].

During September and October 1996, subjective testing was therefore carried out at the BBC Research and Development Department at Kingswood Warren, UK and at NHK Science and Technical Research Labs, Tokyo, Japan.

This report describes, in detail, the various stages of these formal tests and presents the results obtained.

1.2 Test methodology

The methodology for these tests was based extensively on the ITU-R Recommendation BS-1116, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems" [17]. This Recommendation had been prepared specifically to highlight any deficiencies of a sound system. The rationale behind this is that subjective assessments are completed in a matter of hours (per listener), but a consumer of audio hardware and systems is going to be exposed to the resulting quality for many years. It is essential, therefore, from all points of view, to ensure that the tests reveal what the consumer will ultimately find out for him/herself. For this reason, techniques by which the sensitivity of the listeners are maximised (training sessions, etc.,) were included in the Recommendation and have been applied in these tests.

[1] In this context, Backwards Compatibility relates to compatibility with MPEG-1 Audio, IS 11172-3

1.3 Time schedule

A period of eight and a half weeks was available for these tests from receipt of the test material to submission of this report to MPEG. Of this time, four weeks was required for the listening tests themselves.

The time schedule for the tests was as follows:

Date	Activity
28 August 1996	Deadline for receipt of codec submissions at BBC R&D Department
28 August - 2 September	Preparation of all submitted excerpts for auditioning by Selection Panel
2 September - 6 September	Selection panel audition all items and select the excerpts for the test.
9 September - 13 September	Preparation of selected excerpts for the tests
16 September - 11 October	Listening tests in progress
14 October - 28 October	Statistical analysis and Report drafting
28 October 1996	Submission of Report to MPEG

2. Codecs under test

2.1 Codecs proposed for test

The multichannel audio systems under test are all five channel systems with 3 front channels/loudspeakers L, C, R, and 2 surround channels / loudspeakers LS and RS. (The tests were carried out without accompanying pictures.)

The main purpose of these tests was to characterise the performance of the MPEG-2 NBC coding algorithm but additional codecs were also sought to provide comparisons to other currently used multichannel audio coding techniques. Subsequent to the discussions at the July 1996 MPEG meeting, the following codecs were planned to be included:

- MPEG-2 NBC at 320 kbit/s
- MPEG-2 NBC low-complexity version at 320 kbit/s
- Dolby AC-3 in the range 320 to 384 kbit/s
- MPEG-2 Layer II at 640 kbit/s in a backwards compatible mode.

The MPEG-2 Layer II BC codec was included to provide a link back to the results of previous subjective tests undertaken in the RACE dTTb project [4] and to provide justification for the existence of the MPEG-2 NBC codec. It was recognised that this MPEG-2 Layer II BC codec implementation might not reflect the current level of performance of the MPEG-2 Layer II BC codec but, as a reasonable approximation, it could be a guide to the relative improvement in performance to be expected from the MPEG-2 NBC coding techniques. This common element would also be beneficial in comparing the relative performance of these codecs to those in previous tests and would help validate the test procedures themselves. For this reason, precisely the same software implementation of the MPEG-2 Layer II BC codec was used in both these and the earlier RACE dTTb tests.

The Dolby AC-3 codec was to be included, at the request of industry representatives at the July 1996 MPEG meeting in Tampere, in order to provide a comparison between the newly developed MPEG-2 NBC codec and an established non-backwards compatible coding technique.

As will be explained later, the actual codecs in the tests differed from those listed above.

As each codec was implemented in software, the following package of four elements was required for each submission:

- All of the available audio excerpts (94 items), encoded and decoded at the bitrate to be used in the test. These were to be available both as encoded bitstreams and decoded audio files.
- The same test material encoded at a lower bitrate, again both as encoded bitstreams and audio files.
- The encoder software.
- The decoder software.

The lower bitrate versions were primarily for use in the listener training sessions as described in Section 4.2. However, they also allowed the selection panel to work more quickly through the test excerpts and eliminate those which did not invoke artefacts. The encoded bitstreams and the encoder and decoder software were requested to allow independent verification of the coding and decoding processes and to allow test items to be regenerated, should this turn out to be necessary due to files or tapes being corrupted.

The following table summarises the codecs supplied for these tests.

Codec	**bitrate (kbit/s)**	**Supplied by**
MPEG-2 NBC	320	FhG, Dolby, Lucent, AT&T, University of Hannover
MPEG-2 NBC	256	
MPEG-2 NBC	224	
MPEG-2 NBC - low complexity	320	
Dolby AC-3	640	Dolby Labs.
Dolby AC-3	512	Dolby Labs.
MPEG-2 Layer II	640	Philips
MPEG-2 Layer II	512	Philips

As the MPEG-2 Layer II BC codec had already been submitted in a similar way for the RACE dTTb tests [4], the encoding of the test excerpts using this codec was performed at the BBC using this same software. This ensured that the excerpts used in these tests were identical to those used in the dTTb tests. In this case, Philips was also provided with the decoded material to allow them to verify that the submission was as they intended.

2.2 Codec verification

In order to eliminate the possibility of the codecs being 'tuned' for each test excerpt (or within each test excerpt), each codec was submitted to an independent site for verification. The verification sites were supplied by the codec developers, with the encoding and decoding software, the reference and decoded versions of the test excerpts, and the encoded bitstreams. Additionally, DMV were supplied, by the BBC, with copies of the reference and NBC decoded versions being used in the test preparations. This provided the verification that the material being used in the actual tests was identical to that submitted to the verification sites.

Both the 640 kbit/s and 512 kbit/s AC-3 bitstreams were replayed through a commercial AC-3 decoder (the Meridian 565) and were decoded by that hardware. There was inadequate time to verify whether the output of this decoder matched the submitted audio files but brief auditioning of a few items revealed no obvious errors.

The bitrates of the MPEG-2 NBC and MPEG-2 Layer II codecs were also verified for a selection of the test excerpts.

Apart from the Dolby AC-3 submission at the 512 kbit/s bitrate (see Section 2.3), all codecs passed the encode/decode verification procedure.

The table below summarises the codecs supplied for the tests and the verification sites for each.

Codec and bitrate (kbit/s)	Encoding/decoding verification by:	Bitrate verification by:	Verified bitrate (kbit/s)
MPEG-2 NBC (320)	DMV, University of Hanover and Five Bats	Five Bats	320
MPEG-2 NBC (256)			256
MPEG-2 NBC (224)			224
MPEG-2 NBC - low complexity (320)			320
Dolby AC-3 (640)	DMV, Five Bats, BBC	Not evaluated	Not evaluated
Dolby AC-3 (512)	Verification not possible (see section 2.3)	Not evaluated	Not evaluated
MPEG-2 Layer II (640)	Deutsche Telekom and Deutsche Thomson Brandt	Deutsche Telekom	640
MPEG-2 Layer II (512)			512

2.3 The Dolby AC-3 submission.

At the Tampere meeting of MPEG [18], the Audio Subgroup decided to concentrate on the assessment of codecs working at a bitrate in the region of 320 to 384 kbit/s. The only exception to this was to have been the reference to earlier audio codec assessments provided by the 1995 MPEG-2 Layer II codec at 640 kbit/s. Dolby Laboratories were invited to participate with their AC-3 codec on this basis.

Dolby Laboratories announced their decision to participate and advised, on 27 August, that they were submitting the AC-3 codec at a bit rate of 640 kbit/s for testing and 512 kbit/s for training.

Unfortunately, the submission of the AC-3 material at the 640 kbit/s and 512 kbit/s bitrates by Dolby Laboratories did not fulfil the needs of the industry request for a test of the commercial coder (typically 384 kbit/s). Neither did this agree with the requested bitrate range of the formal listening test ad hoc group for use in a comparison to the MPEG-2 NBC coder being characterised at 320 kbit/s.

While a great deal of additional effort was expended at the BBC test site attempting to make use of the submitted AC-3 material, many difficulties were encountered. The verification of the material at 512 kbit/s could not be carried out[2], any consequential substitution of replacement AC-3 material could have invalidated the blind nature of the selection panel process and no time was available to have the sequences re-encoded at a bitrate appropriate for the test. This led to the ultimate admission that the needs for including AC-3 in the test could not be addressed with the submitted material. It was regrettably concluded that the submitted AC-3 codec could not be included in the tests and that, in the time remaining, it would be impossible to substitute a revised version. This decision was communicated to the MPEG Audio Subgroup by email from the Chairman on 6 September 1996.

After this decision was taken, it was necessary to consider whether the selection panel's choice of the ten critical test items was still appropriate for the remaining codecs. As can be seen in their Report, the selection panel had identified coding artefacts in all the codecs and so this was difficult as neither the identity of the codecs nor the above decision could be revealed to the panel. However, after general discussions with them, it appeared that the ten test items still represented a balanced selection for the remaining codecs.

2.4 Codecs tested

Following the decision that the AC-3 codec could not be included in the tests and the feedback from the selection panel indicating that the quality of the codecs was high, it was decided that the MPEG-2 NBC codec at the lower bitrate of 256 kbit/s should also be included in the tests. This would give an opportunity to explore what could possibly be the lower region of the operating range for this codec.

The following codecs were therefore included in the tests:

- MPEG-2 NBC at 256 kbit/s
- MPEG-2 NBC at 320 kbit/s
- MPEG-2 NBC low-complexity version at 320 kbit/s
- MPEG-2 Layer II at 640 kbit/s in a backwards compatible mode.

2.5 Status of features used in the MPEG-2 NBC codecs

The main profile NBC multichannel encoder / decoder is compliant with the MPEG-2 NBC Committee Draft. The main profile has the features MS stereo, intensity stereo, NEC lossless coding, prediction, temporal noise shaping and dynamic switching of window shape variously available for the different modes. Not all of these features were necessarily active in each of the embodiments.

[2] Dolby Laboratories advised that the 512 kbit/s training bitstreams had been created on several different computer platforms and could therefore not be verified. Furthermore, the encoder software, which was necessary to perform the verification, had not been supplied.

All NBC options used the same decoder, which is fully compliant with the MPEG-2 NBC Committee Draft syntax.

The encoders were set to have only the following features active for these tests:

1. MPEG-2 NBC at 320 kbit/s.

This coder used very conservative parameters. Prediction and temporal noise shaping were turned on, but most additional features were switched off to provide the smallest possible change to the RM4 version evaluated earlier in the project [19].

2. MPEG-2 NBC at 256 kbit/s

The NBC coder at 256 kbit/s used a combination of prediction, MS stereo coding, temporal noise shaping and dynamic switching of window shape.

3. MPEG-2 NBC Low Complexity at 320 kbit/s

Complexity reduction centres on omitting prediction from NBC: thus prediction was turned off. MS stereo coding, temporal noise shaping (of a lower order than for main profile NBC) and dynamic switching of window shape were activated.

3. Test material

3.1 Call for test excerpts

A call for suitable five channel test excerpts with a duration of about 20 seconds was distributed in March 1996 to MPEG members and others working in this field [20]. In total, 94 test excerpts were provided, comprising 66 items made available for earlier tests and 28 new items offered by Deutsche Telekom TZD, Decca Recording, the University of Surrey, NHK and the BBC.

The 28 new test excerpts were matched in level to the earlier excerpts at BBC R&D and converted into individual files (in AIFF). Together with the earlier 66 items, they were then copied to Exabyte or Data DAT for distribution to the codec developers during July 1996. Altogether, the 94 excerpts amounted to just over 1 Gbyte of data.

3.2 Selection of test excerpts for the test

A selection panel was established primarily to identify the ten critical excerpts to be used for the tests. Their tasks are detailed in [16] but, briefly, were:

- to determine, from the 94 excerpts available and using all the codecs, the ten most critical items, whilst avoiding material of a similar nature
- to recommend which of the selected ten test excerpts should be used for listener training
- to ensure that the range of selected codec/item combinations included a number of test items which were likely to invoke grades in the region of 3 to 3.5 on the impairment scale
- to identify any codec/bitrate combinations which consistently offer poor quality
- to offer advice concerning the tests, having auditioned the test excerpts

The selection panel consisted of:

- Thomas Buchholz, Deutsche Telekom, TZD-Berlin,

- Kazuho Ono, NHK, Science and Technical Research Labs, Tokyo,
- Andrew McParland, BBC Research and Development Department,
- John Fletcher[3], BBC Research and Development Department.

The panel carried out their work in Listening Room 2 (the same room subsequently used for the formal tests) at the BBC's R&D Department, Kingswood Warren.

Immediately prior to the selection panel meeting, a pre-selection was performed on some of the test excerpts by auditioning those low bitrate versions which by then had been prepared. This proved invaluable in reducing the number of items which needed to be auditioned in detail by the full selection panel.

3.3 Preparation of test excerpts for the selection panel

In total, seven versions of all the 94 test excerpts were available for the selection panel (high and low bitrate versions of NBC, Layer II BC and AC-3, and one version for NBC-low complexity). In order to conceal their identities, each codec, regardless of bitrate, was assigned an identification letter A, B, C or D for the selection panel work. Apart from the Layer II BC versions (which had been coded at the BBC), the audio files where received on Exabyte tape, converted to AIFF and loaded, via an Ethernet link, onto a Sonic Solutions Audio Editor. (This transpired to be a very time-consuming operation because of the large quantity of data involved.)

Although it had been planned to provide synchronised coded and reference recordings for the selection panel, this proved to be too ambitious an undertaking in the two days which remained available for material preparation, once all the submitted material had been received. Instead, a single Tascam DA88 recorder was used and the material recorded in the order Reference, codec A, codec B, Reference, codec C, codec D, Reference for each of the 94 items in turn. Spoken announcements ("Reference", "A", etc.) preceded each version. The generation of edit lists to create these tapes was almost entirely automated, otherwise this would not have been feasible in the time available.

Two sets of Tascam DA88 tapes were then created, one at the lower bitrate and the second at the higher bitrate (the low-complexity NBC version was included at the same bitrate in both). In total, ten 2-hour Tascam tapes were required for these recordings.

During their work, the selection panel were able to control the replay of the tape using custom software which offered a menu of items on a screen. Simple keyboard commands allowed items to be selected and replayed as necessary in any order by the selection panel themselves.

3.4 Results from the Selection panel

All the different steps of the selection panel work, as well as the complete list of the 94 test excerpts and descriptions of coding artefacts and codec characteristics, can be found in the selection panel report which is given in Annex A.

The table below lists the ten test excerpts recommended by the selection panel for the tests.

[3] John Fletcher was invited to join the selection panel after carrying out a significant part of the pre-selection process with Andrew McParland, on the day before the full panel met.

No.	Name	Description
1	pitch_pipe	Pitch Pipe
2	harpsichord	Harpsichord
3	triangle	Triangle
4	cast_pan1	Castanets panned across the front, noise in surround
5	elliot1	Female and male speech in a restaurant, chamber music
6	mancini	Orchestra - strings, cymbals, drums, horns
7	station_master1	Male voice with steam-locomotive effects
8	clarinet_theatre	Clarinet in centre front, theatre foyer ambience, rain on windows in surround
9	thalheim1	Piano front left, sax in front right, female voice in centre
10	glock	Glockenspiel and timpani

Of these items, the panel recommended harpsichord, triangle, Mancini and Thalheim for use as the main items in the training session.

3.5 Low anchor presentations

As the selection panel proceeded with their work and began to report back their findings, it became apparent that the tests themselves may not include a sufficient number of low anchor presentations which are essential in proving the test as a whole [16].

A request was therefore made to FhG to see if the test material was available, encoded through either NBC implementation, at a yet lower bitrate. FhG were able to provide the material processed by the NBC codec at 224 kbit/s but this had not been checked by them in detail and so it could not be included in the formal tests. Nevertheless, the material was received and auditioned by the selection panel, with a view to using some excerpts as lower anchor items or as training material (particularly with the NBC codec now being tested at 256 kbit/s). The conclusion of this assessment was that the material was not suitable for the low anchor presentations but would be useful in the training sessions.

In order to check the Selection Panel's view that few low anchor presentations would be in the test, a brief listening test was arranged with two experienced listeners, not directly associated with this work. A blind test was set up with the four most critical test excerpts and using codecs which would be in the formal tests. The results of this showed, in one case accurate identification but reasonably high grades and in the other case, more errors in identification but lower grades. As no distinct pattern emerged, it was concluded that it would be wise to include other low anchor presentations in the tests, otherwise the validity of the test results could have been jeopardised.

Versions of four of the test items, which were likely to give mid-range quality, were therefore identified from the results of the MPEG '94 tests [1] as suitable low-anchors. These were chosen on the basis of the mean grades which the items were awarded in those tests: the codec identities and bitrates were not important in this choice. Using this criterion, the items chosen were: Harpsichord through the MPAC codec (at 320 kbit/s), Mancini through the Layer II codec (at 320 kbit/s) and Pitch

pipe and Triangle, both through the AC-3 codec (at 320 kbit/s). Although it may have been beneficial to include yet more low-anchor presentations, there was concern that the durations of the test sessions would then be lengthened unacceptably for the listeners.

With the agreement of both test sites and the Chairman of the MPEG Audio Subgroup, these four items were added to the tests to make a total of 44 presentations (4 codecs * 10 excerpts plus these four low-anchor items). As the inclusion of these low-anchor items was an aspect which could have influenced the grades awarded by listeners in judging the presentations, it was not communicated to the selection panel or the MPEG formal test group.

3.6 Preparation of excerpts for the test

The ten selected test excerpts, in their original and decoded versions (for each codec which was available) were copied to Exabyte tape for the NHK test site. Difference files (i.e. reference - coded version) were also included for use in the training sessions.

In this preparation process, the codecs were randomly assigned the identities V, W, X or Y. These identities were used for the remainder of the test preparations at both test sites.

4. Experimental design

The test design followed the ITU-R Recommendation BS-1116 [17] and listeners were asked to judge the single, all-embracing attribute “Basic Audio Quality” as proposed in that Recommendation.

According to the test specification [16], four weeks were available to assess the quality of the four codecs with at least 20 listeners at each test site.

At both test centres, three listeners participated every two days. The first half-day was used for training as a group, with the remaining one and half days available for the grading phase.

4.1 Test procedure

The tests used the “triple stimulus/hidden reference/double blind” method. The listener could switch freely between the presentations “Reference”, “A” and “B”, where “A” and “B” are the processed version and the hidden reference, randomly allocated from one trial to the next. (The allocation was known neither to the listener nor to the test supervisor, hence the term “double blind”.)

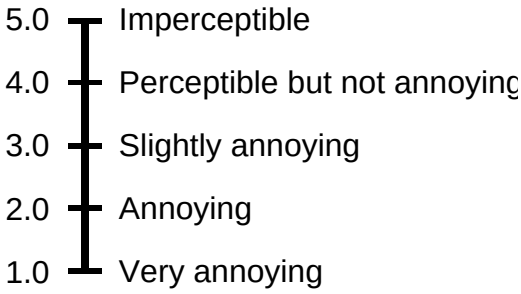
The tests were organised with an initial training phase, for all listeners involved during the session, and a grading phase, in which the listeners individually carried out the assessments.

The grading phase was carried out by the listeners individually in 8 sessions each of which took about 25 to 30 minutes to complete. In each trial, the listener heard three versions, labelled on the computer screen as “Ref”, “A” and “B”, and could switch freely between them at any time. The listening level of all the material under test was fixed. Each test excerpt could be repeated as often as the listener wished. The

listener was asked to judge the “Basic Audio Quality” of the “A” and “B” versions in each trial. This attribute is related to any and all differences between the reference and the coded/decoded programme excerpt.

Any difference between the reference and the coded/decoded programme excerpt was to be considered as an impairment. Anything that the listener detected as a difference had to be included in their overall rating.

Each listener graded the perceived differences using the following grading scale (the ITU-R five point impairment scale) :



The grading scale was to be considered as a continuous equal interval scale but with descriptions at five “anchor points” to indicate specific values.

The listeners were asked to input their grades to an accuracy of one decimal place.

At least one grade of “5.0” had to be given for each trial, since one of “A” or “B” was the hidden reference.

The order of the test presentations and the position of the hidden reference was randomised for each test listener. Therefore, any comments made by one listener during the test phase would not be relevant to the perceptions of the other listeners.

4.2 Training

The morning of the first test day was used for a joint training session involving the three listeners for that two-day session. This allowed them to become familiar with the test procedure, assist each other in identifying coding artefacts and generally to become more experienced listeners. The listeners were guided during this training by a test supervisor.

It had been agreed, in advance, that the training could make use of bitrates lower than those used in the tests. This would make the impairments clearer for listeners, particularly for those who were not familiar with this type of artefact. The test centres, therefore, made use of the lower bitrate material to guide listeners in the early stages of the training session.

To maintain the blind nature of the tests, both the test supervisor and the listeners were unaware of the identities of the codecs and bitrates being used in the training sessions.

Throughout the training and the tests, the listeners were asked not to discuss the grades they would award for audio quality as this was required to be an individual subjective judgement for each of them.

The steps taken in the training phase were similar at both test centres and followed the pattern:

Step 1. An initial impression of the ten test excerpts and coding artefacts was demonstrated by replaying the reference version and a coded version of each item. In this case the coded version was the MPEG-2 NBC codec operating at 224 kbit/s. This introduction allowed listeners to become familiar with the 5-channel arrangement and also to hear typical coding artefacts.

Step 2. Each of the four training items (a subset of the ten test items), coded with one of the codecs at a lower bitrate (alternating between NBC at 256 kbit/s and Layer II BC at 512 kbit/s) was replayed in reference and coded forms with a short break after each coded presentation for discussions about the perceived artefacts. For each item, once listeners had discussed what they had heard, the difference signal, i.e. the difference between the coded and reference signals, was replayed. It was explained carefully that this would contain elements which would be inaudible in the normal presentation, but, nevertheless, could indicate where it may be possible to hear artefacts and their nature. Once this had been replayed, the reference and coded version of the item were again replayed and further discussion encouraged. This procedure was repeated for each of the four main training items in turn.

Step 3. This step concentrated on each of the four items in turn, presenting each initially at the lower bitrate but through the other codec of the two used in Step 2. As before, the difference signals were then presented followed by the reference and coded versions. This same item was then replayed at the different bitrates and through the other codec of the pair. This process exposed the listeners to the range of coding artefacts which would be encountered in the tests.

Step 4. This step repeated Step 1 but for the remaining six test items, i.e. those not scrutinised during Steps 2 and 3. This was to allow the listeners to consider the artefacts which may be audible in these remaining items, after completing the detailed auditioning of the four items.

Because of the limited time available and the need to adapt the training to the listeners present, some of the elements of the training were omitted at times. For example, Step 2 was sometimes omitted because these presentations would be repeated in Step 3. Also, not all the difference signals were replayed.

After completing the above training, the listeners each carried out a ‘mini-test’. This used the four main training items arranged as a randomised Ref/A/B test. The purpose of this was two-fold: to allow practice with the test control system and also to accustomise each listener to individual listening (which is psychologically harder than collective listening). Each listener was allowed about ten minutes to do this test and was advised that the results were not important and would not be used.

4.3 Listening panel

According to the requirements of ITU-R Recommendation BS-1116 [17], the listening panel should consist of at least 20 expert listeners at each test site in order to get statistically meaningful results.

Requests were therefore made to various groups involved in audio work. In the end, more than 20 expert listeners were available at each of the two test sites.

In advance of their participation in the tests, further information was sent to each listener about the tests including details of the test method, test procedure and time schedule. This information is included as Annex B.

4.4 Listening conditions and test equipment

ITU-R has defined specific requirements for the listening conditions to ensure comparable and reliable results of subjective assessments of sound systems [17].

This covers:

- the acoustical characteristics of the listening room and the sound field therein,
- the arrangement of the monitoring loudspeakers in the listening room,
- the location of the listening positions for the test.

The listening rooms used at both of the test sites fulfil the majority of the corresponding requirements.

The listening arrangement as given in [17] was used at both test sites, with the listener in the “Centre” or “Reference Listening Position”.

The most important technical parameters and acoustic characteristics which describe the sound field affecting the listener at the Reference listening position, namely,

- geometric properties of the listening room
- the operational room response curve
- the frequency response of the reverberation time
- the listening arrangement used
- the background noise level
- other technical parameters

are presented in Annex C for the BBC and Annex D for NHK.

The loudspeakers used at the two test sites were of different types, but, in each case, high quality studio monitors were used. All the loudspeakers used fulfil the requirements of the ITU-R Recommendation.

In accordance with [17], the listening level at the Reference listening position was adjusted to an SPL of 78 dB(A) for each loudspeaker, by means of a pink noise signal with the same RMS value as a 1 kHz tone at -18 dBFS. The maximum SPL of the test excerpts then reached about 75 to 95 dB(A), depending on the programme content and the matching of the perceived loudness of each excerpt.

5. Test arrangements at each test centre

5.1 Arrangements at the BBC

5.1.1 Listening room.

Listening Room 2 at BBC Research and Development Department, Kingswood Warren, was used for these tests. This room was used for the previous multichannel tests: the MPEG tests in 1994 [1] and the Race dTTb tests in 1996 [4]. The characteristics of this room are given in Annex C. Although slightly smaller than that recommended in the ITU-R Recommendation [17] for multichannel sound tests, in most other aspects the requirements are met.

5.1.2 Test Equipment

The playback system at the BBC used two Tascam DA88 digital audio eight-track recorders to replay the reference and coded recordings as shown in Annex C.

Five tracks on each machine were used, with one DA88 replaying the reference recording and the second replaying the coded version in sample accurate synchronisation.

A Tascam to AES/EBU and a Tascam to Yamaha digital format converter were used to feed the signals to the main and monitor inputs of a Yamaha DMC1000 mixer which performed the switching between reference and coded signals. The digital outputs of the DMC1000 were fed to three Prism DA-1 DACs to give the five analogue output signals. These were passed through 1/3 octave graphic equalisers to the power amplifiers. The replay level was set by the gain of the power amplifiers.

Control of the tests was carried out using custom software developed for conducting subjective tests. The screen display used during a test session is shown in Annex C. In addition to providing the selection of Reference, A and B presentations and the entering of grades, the software also provides control of the audio replay machines.

Although the software can be controlled by using a mouse or keyboard, all listeners preferred to use the hand-held keypad shown in Annex C, which connects to the control system via an RS232 interface. It offers all the functions necessary to conduct the tests and enter grades.

5.1.3 Preparation of test excerpts

The ten test excerpts for the four codecs, plus the four low anchor items, were compiled into eight test blocks, each containing either five or six test excerpts but with codecs and test excerpts in a randomised order. This preparation was carried out using a Sonic Solutions Audio Editor with the edit decision lists being created automatically by custom software which was fed with the required randomisation sequence. Each test excerpt was replayed, sample aligned with its corresponding reference, with nine repetitions and recorded on to two Tascam DA88 tapes.

5.1.4 Listening panel

The listeners participating at the BBC were all professionally involved in audio work. The majority of listeners had a background of sound production for television or radio, whilst the remaining listeners were involved in audio engineering. The majority

had already gained experience of low bit-rate coding and had participated in earlier tests on two-channel or multichannel systems.

In total, 32 listeners participated in the tests at the BBC. The list of the test participants and their affiliations is given in Annex E.

5.2 Arrangements at NHK

5.2.1 Listening room.

Listening Room B268 at NHK Science and Technical Research Laboratories, Tokyo Japan, was used for these tests. The characteristics of this room are given in Annex D. The requirements in the ITU-R Recommendation [17] for multichannel sound tests are met in most respects.

5.2.2 Test Equipment

The playback system at the NHK used a Sonic Studio DAW (Digital Audio Workstation) running on an Apple personal computer, (a Power Macintosh 8100/100AV), to replay the reference and coded recordings. This is shown in Annex D.

Ten tracks on the DAW were used with five tracks replaying reference recording and the other five tracks replaying the coded version.

The mixing desk which is included in the Sonic Studio DAW was used to perform the switching between reference and coded signals. From the DAW, the selected five audio channels were output in AES/EBU format and then fed to three YAMAHA DA-2X DACs to give the five analogue output signals. These then passed through 1/3 octave graphic equalisers (Yamaha DEQ5) to the power amplifiers. Initially, the equalisers were used to adjust the operational frequency response. However, they were later bypassed because the reproduced sounds without the equalisers were felt to be better than with them. This decision was taken before any formal tests had started. The observation was also made (by Mr Ono, a member of the selection panel) that the sound quality without equalisation was then more similar to that which he had heard at the BBC tests site.

Replay level was set by the attenuators and the gain of the power amplifiers.

Control of the tests was carried out using custom software developed for conducting subjective tests on another Apple personal computer, 2Vi. Randomisation of the tests was also conducted with this software. The software also provides control of the DAW, which is implemented through the network between the two personal computers. In addition, the software also provides the selection of Reference, A and B presentations and allows the inputting and the recording of grades. The screen display used during a test session is shown in Annex D.

The software was controlled by the hand-held keypad shown in Annex D, which connects to the control system via the ADB (Apple Desktop Bus) interface. It offers all the functions necessary to conduct the tests and enter grades.

The test facility allowed different presentation orders for each listener and the listeners worked through their own sequence, in sessions of typically 25 to 30 minutes, stopping and resuming within the sequence as appropriate.

5.2.3 Preparation of test excerpts

The selected test items were received by NHK on Exabyte tape from the BBC. For the tests, the excerpts were loaded on to the Sonic Studio DAW from which they could be replayed, with the synchronised reference recording, in any order.

5.2.4 Listening panel

The subjects participating at the NHK were all professionally involved in audio work.

In total, 24 subjects participated in the tests at the NHK. The list of the test participants and their affiliations is given in Annex E. 19 listeners were involved in audio engineering whilst 5 listeners had a background of sound production for television or radio.

6. Statistical analysis and results

6.1 General

The aim of the analysis was to answer the following questions, which were itemised in the test specification [16]:

1. Are the listeners' results reliable, i.e. distinguishable from random votes?
2. Does the test methodology allow meaningful conclusions to be drawn from these results?
3. Is there any distinction between the two test sites?
4. Is the performance of NBC at the default bitrate *[320 kbit/s]* equal to or better than the performance of *[the 1995 version of]* BC Layer II at 640 kbit/s?
5. How does the performance of the codecs vary with programme items?
6. Is the performance of the coding of NBC at the default bitrate *[320 kbit/s]* distinguishable from the original signal?
7. Is the performance of NBC at the default bitrate *[320 kbit/s]* achieving 'indistinguishable quality' in the EBU definition [21] of that phrase?
8. What is the relative ranking of the codecs tested?
9. Are there any other features from the data that should be reported?

This Section presents the key steps in the statistical analysis and the corresponding results obtained to answer these questions. Some further aspects of the statistical analysis are discussed in Annex F, whilst Annex G contains the detailed numerical results from these analyses.

For the analysis, the raw data from each test site was unscrambled at the test site prior to being forwarded to the Mathematics and Computing Science Department at the University of Surrey.

'Diffgrades' are used throughout the analysis; these are calculated, from each trial, as the grade awarded to the coded version minus the grade awarded to the reference. The table below shows the 5-grade impairment scale used by the subjects and the equivalent diffgrades corresponding to the ITU-R Recommendation BS-1116 [17]. (This table assumes that the listener has graded the coded version rather than the hidden reference).

IMPAIRMENT	Grade	Corresponding diffgrade
imperceptible	5.0	0.0
perceptible, but not annoying	4.0	-1.0
slightly annoying	3.0	-2.0
annoying	2.0	-3.0
very annoying	1.0	-4.0

Negative diffgrades indicate that the subject correctly identified the coded version whereas a positive diffgrade indicates that the subject failed to identify the coded version. As can be seen from the table above, a diffgrade closer to zero means that the subject judged the audio quality to be higher.

6.2 Post-screening to assess listener reliability

As suggested in the ITU-R Recommendation BS-1116 [17], a post-screening of all the listeners was carried out by using a one-sided t-test at the significance level, α=0.05. The probability of accepting a subject who, on average, was unable to detect the coded version, is then 0.05 at most. In addition to this procedure, a Wilcoxon test was also applied to assess reliability.

The results from the t- and Wilcoxon tests for the listeners attending the BBC test site are tabulated below. To be accepted by either test, a listener must achieve a probability level (the final two columns in the table) below 0.10, corresponding to a 5% threshold for a one-sided test. It can be seen that, for the t-test, 9 of these listeners (the shaded entries) exceed this value and have therefore been removed from the subsequent analysis. This is confirmed by the Wilcoxon test for all but one listener who would be included by this latter test (see Annex F for a discussion on this aspect). Accordingly, the rest of the analysis has been carried out with the remaining 23 listeners from the BBC site.

BBC Listener	Mean	St. deviation	t-test	Wilcoxon
1	-0.5455	0.888	0.000	0.0005
2	-1.7159	1.322	0.000	0.0000
3	-0.5841	0.929	0.000	0.0002
4	-1.4318	1.701	0.000	0.0000
5	-0.8455	1.102	0.000	0.0000
6	-1.7227	1.404	0.000	0.0000
7	-1.4136	1.336	0.000	0.0000
8	-0.5773	1.186	0.002	0.0035
9	-0.6295	1.251	0.002	0.0042
10	-0.3000	0.635	0.003	0.0055
11	-0.2364	0.518	0.004	0.0090
12	-0.2568	0.582	0.005	0.0060
13	-0.3545	0.820	0.006	0.0023
14	-0.3273	0.833	0.013	0.0149
15	-0.6364	1.775	0.022	0.0155
16	-0.2364	0.666	0.023	0.0256
17	-0.2795	0.821	0.029	0.0132
18	-0.3773	1.106	0.029	0.0386
19	-0.3477	1.027	0.030	0.0543

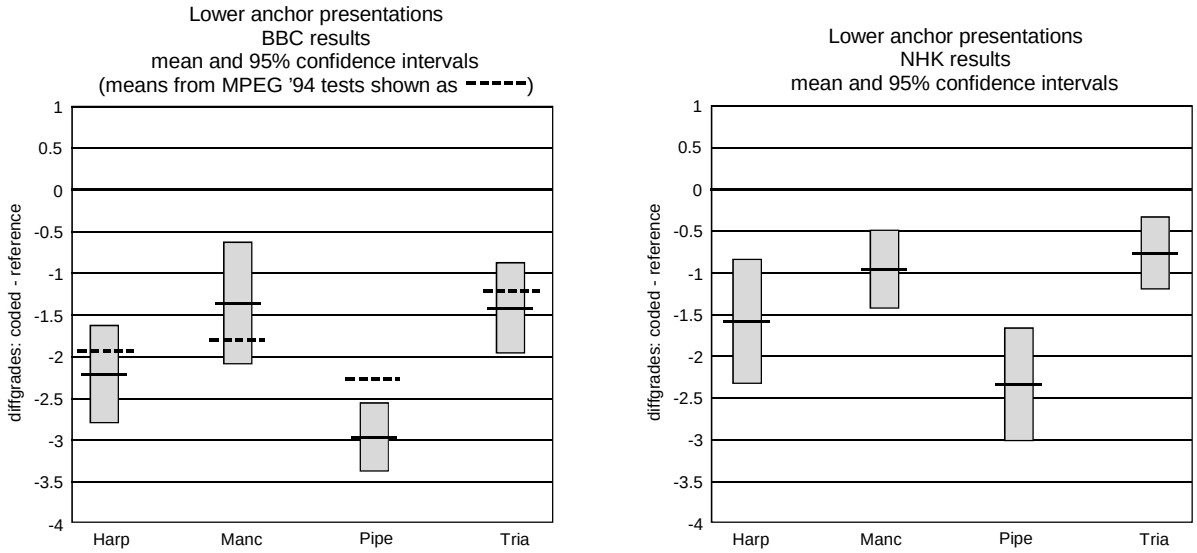
BBC Listener	Mean	St. deviation	t-test	Wilcoxon
20	-0.6591	2.022	0.036	0.0590
21	-0.2773	0.933	0.055	0.0790
22	-0.2500	0.866	0.062	0.0457
23	-0.1682	0.633	0.085	0.0428
24	-0.3068	1.464	0.172	0.0910
25	-0.1523	0.769	0.196	0.3380
26	-0.1068	0.840	0.404	0.5214
27	-0.1818	1.556	0.442	0.4372
28	-0.0614	0.584	0.489	0.5408
29	-0.0341	0.395	0.570	0.8421
30	-0.1000	1.535	0.668	0.5714
31	0.0341	0.760	0.767	0.5478
32	0.0227	1.732	0.931	0.7766

In the same way, the results from the t- and Wilcoxon tests for the 24 listeners attending the NHK test site are tabulated below. It can be seen that 8 of these listeners exceeded the t-test probability threshold (0.10) and must be therefore be removed from the subsequent analysis. This is confirmed by the Wilcoxon test. Accordingly, the rest of the analysis has been carried out with the remaining 16 listeners from the NHK site.

NHK Listener	Mean	St. deviation	t-test	Wilcoxon
1	-0.4409	0.713	0.000	0.0000
2	-0.5727	0.669	0.000	0.0000
3	-1.1023	1.292	0.000	0.0000
4	-0.3636	0.650	0.001	0.0011
5	-0.3955	0.838	0.003	0.0058
6	-0.4659	0.997	0.003	0.0001
7	-0.2955	0.668	0.005	0.0056
8	-0.8864	2.060	0.007	0.0088
9	-0.2114	0.544	0.013	0.0034
10	-0.2955	0.859	0.028	0.0613
11	-0.0773	0.238	0.037	0.0527
12	-0.1909	0.601	0.041	0.0669
13	-0.0727	0.241	0.051	0.0751
14	-0.3750	1.267	0.056	0.0429
15	-0.1159	0.442	0.075	0.0723
16	-0.3955	1.439	0.075	0.0405
17	-0.1932	0.891	0.157	0.1423
18	-0.1868	1.175	0.299	0.3459
19	-0.3250	2.088	0.308	0.3782
20	-0.0409	0.279	0.336	0.4956
21	0.0955	0.827	0.448	0.5568
22	-0.0795	1.007	0.603	0.5730
23	0.0068	0.481	0.925	0.6872
24	0.0114	1.445	0.959	0.8019

6.3 Results for the low-anchor presentations

As the additional four low anchor presentations originated from different codecs, and gave data for only 4 of the ten test excerpts, their data was removed from the remainder of the analysis, as having only this partial data available would have complicated the subsequent analysis significantly. However, the results for these presentations were calculated separately and the results are shown below. For the BBC results, the corresponding mean diffgrades from the MPEG '94 tests [1] are also shown.



From the above BBC results for the low-anchor presentations, it can be seen that the means from the previous tests for three of the four excerpts lie within the 95% confidence intervals produced by these tests.

For the fourth excerpt, pitch pipe, the apparent difference was investigated by using a two-sided t-test. This indicated that, at the 95% confidence level, there is a significant difference between the means from each test (p-level: 0.039): in these tests, the pitch pipe item has been marked more critically than in the MPEG '94 tests. Although no clear explanation for this can be given, it should be noted that, in the MPEG '94 tests, this particular presentation of pitch pipe gave the highest quality for the pitch pipe excerpt, whereas, in contrast, it has the lowest mean grade for all 44 presentations in this test. These overall differences in codec quality between the tests may have led to subjective differences in grading, with the mean diffgrade from the MPEG '94 tests being improved and/or the mean diffgrade from this test being reduced.

These results show that the test arrangements, *i.e. the listening conditions and listeners together*, at both test sites, were able to reveal artefacts below the level of grade 4.0: thus the validity of the test arrangements is confirmed. (As already discussed, further results in this region would have been beneficial in this assessment but the listening time would then have been increased unacceptably.)

6.4 Summary of all effects: Analysis of Variance

Using the data from the listeners who had passed the test for reliability (but excluding the data for the low anchors), a three-way ANOVA was performed with main effects of "Site", "Codec" and "Item". This was primarily to identify if the data from the two test sites could be combined for the remaining analysis. In order to do this, the

three-way ANOVA must show that the effect of “Site” was not significant both as a main effect and in the interactions with the other effects.

The table from this ANOVA is shown below.

Source of Variation	Sum of Squares	DF	Mean Square	F - ratio	P - level
Main Effects	82.534	13	6.349	6.549	0.00
SITE	13.626	1	13.626	14.056	0.000
CODEC	18.908	3	6.303	6.501	0.000
ITEM	49.999	9	5.555	5.731	0.000
2-Way Interactions	73.845	39	1.893	1.953	0.000
SITE CODEC	6.133	3	2.044	2.109	0.097
SITE ITEM	10.299	9	1.144	1.180	0.303
CODEC ITEM	57.412	27	2.126	2.193	0.000
3-Way Interactions	26.131	27	0.968	0.998	0.467
SITE CODEC ITEM	26.131	27	0.968	0.998	0.467
Explained	196.052	79	2.482	2.560	0.000
Residual	1434.777	1480	0.969		
Total	1630.829	1559	1.046		

Whenever the observed significance levels (values of the last column, p-level) are less than 0.05, the corresponding effect has a significant influence. Thus, in this case, the main effects of Site, Codec and Item all show a significant influence at the 5% level. Similarly, the interaction effect between Codec and Item is significant.

As the Site effect was shown to be significant, the data from both test sites could not be combined and so the remainder of the analysis was performed separately for the BBC and NHK data and two sets of results are presented.

6.5 Results from the BBC test site

6.5.1 Two-way ANOVA

A two-way fixed-effects ANOVA using the BBC data produced the following results:

Analysis of Variance for diffgrades: BBC Site, 23 listeners.

Source	Sum of squares	Degrees of freedom	Mean Square	F-ratio	P-level
Main Effects	67.977	12	5.665	5.129	0.000
CODEC	22.103	3	7.368	6.670	0.000
ITEM	45.875	9	5.097	4.615	0.000
2-Way Interactions	60.960	27	2.258	2.044	0.001
CODEC ITEM	60.960	27	2.258	2.044	0.001
Explained	128.937	39	3.306	2.993	0.000
Residual	971.974	880	1.105		
Total	1100.911	919	1.198		

From this table, it can be seen that the Codec main effect, the Item main effect, and the Codec*Item interaction were all statistically significant for the BBC data.

The BBC data gives the following mean values for the overall performance of each codec. It must be noted, however, that item-to-item variations may also exist but these may not be apparent in these mean values.

Mean diffgrades: BBC results

Codec	Number of samples	Mean diffgrade
Layer II at 640 kbit/s	230	-0.49
NBC at 256 kbit/s	230	-0.73
NBC at 320 kbit/s	230	-0.34
NBC lc at 320 kbit/s	230	-0.36

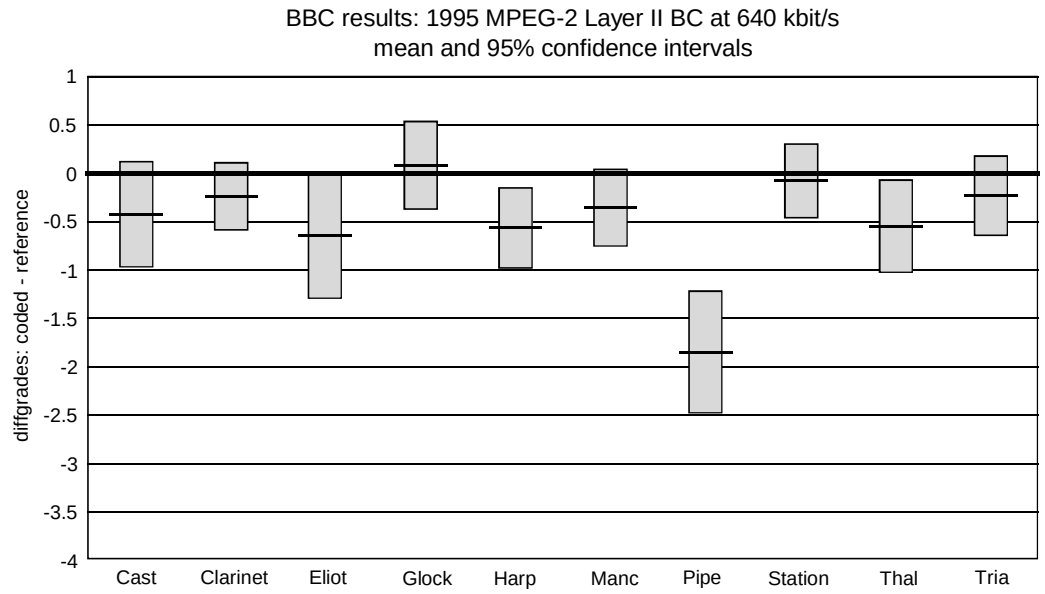
6.5.2 Estimates of means and 95% confidence intervals

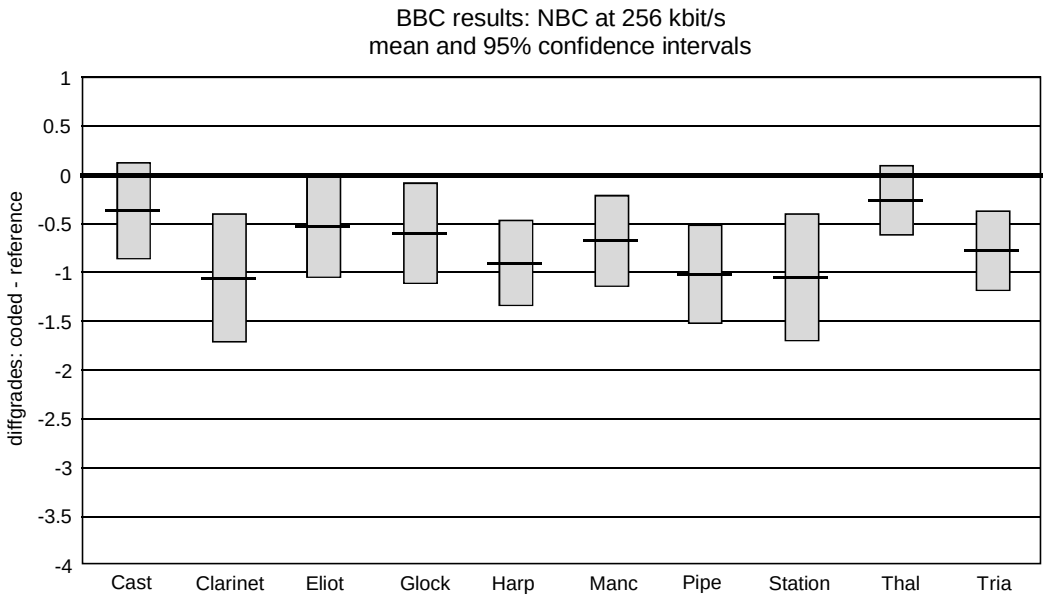
The following four diagrams show the estimated means and two-sided 95% confidence intervals for each codec and excerpt at the BBC site.

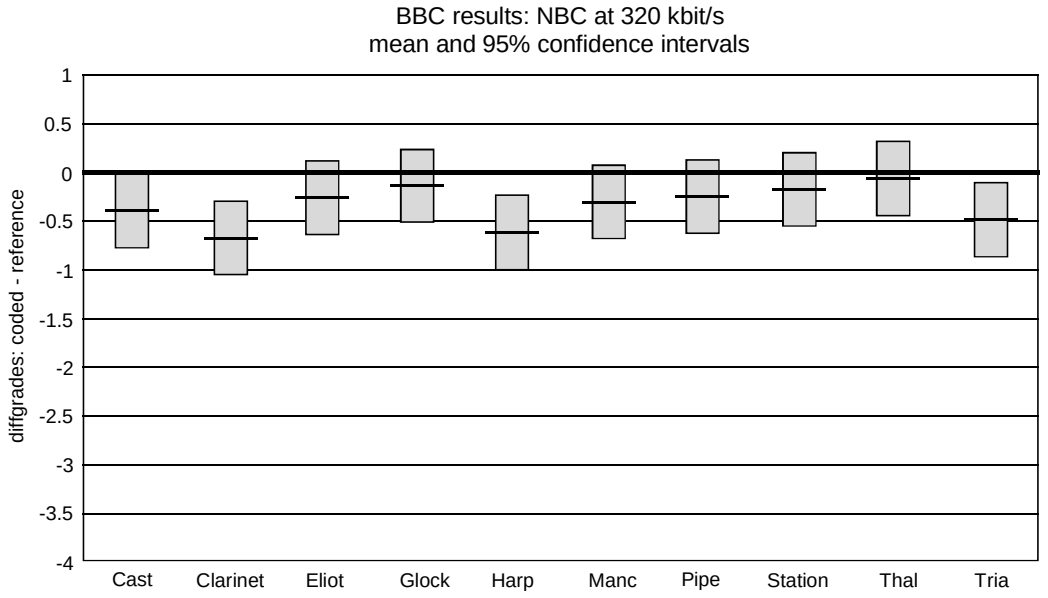
In calculating the confidence intervals, the model assumptions for the ANOVA have been checked by applying Levene's test. Because this test showed that the variances were unequal, the strict assumptions for the ANOVAs for the MPEG-2 Layer II BC codec at 640 kbit/s and MPEG-2 NBC codec at 256 kbit/s were not satisfied. Thus, within the BBC data, there is heterogeneity of variance between items for these two codecs. Therefore, when comparing individual items for these codecs, the error mean square from the ANOVA has not been used. However, the overall ANOVA results will be broadly interpretable, because ANOVA is robust to modest departures from the homogeneity of variances assumption.

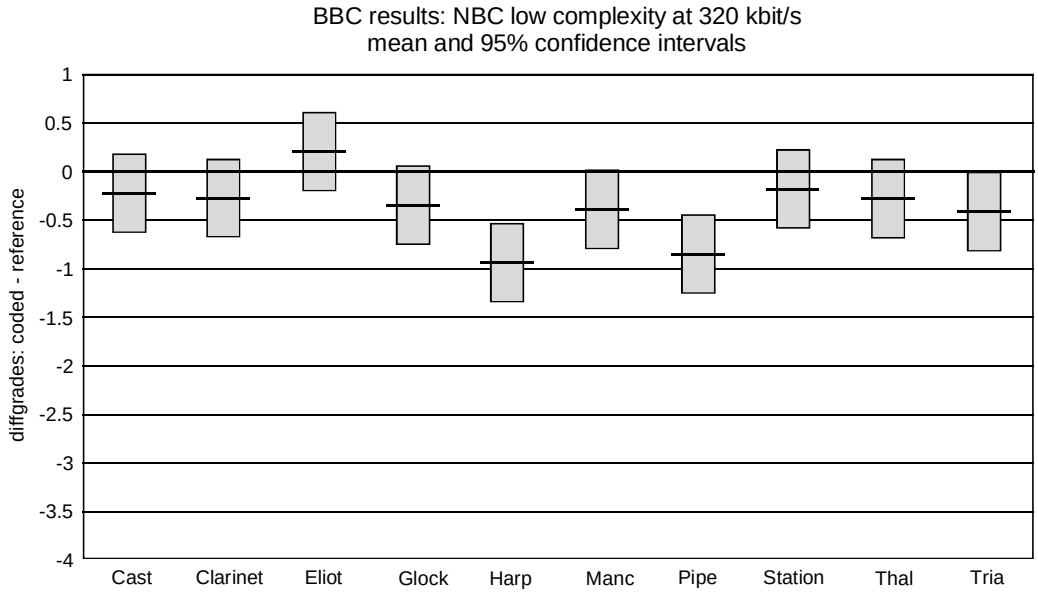
Where the model assumptions are valid (for NBC at 320 kbit/s and NBC low complexity at 320 kbit/s), the confidence intervals have been computed from a one-way ANOVA for each codec. Thus these confidence intervals possess equal length for each of the items. Where the model assumptions are rejected, the confidence intervals are calculated from the individual estimates of the standard deviations.

The results of the Levene's test, the ANOVAs and the tabulated data for these diagrams are given in Annex G.


BBC results: 1995 MPEG-2 Layer II BC at 640 kbit/s
mean and 95% confidence intervals
diffgrades: coded - reference
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
Cast
Clarinet
Eliot
Glock
Harp
Manc
Pipe
Station
Thal
Tria


BBC results: NBC at 256 kbit/s
mean and 95% confidence intervals
diffgrades: coded - reference
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
Cast
Clarinet
Eliot
Glock
Harp
Manc
Pipe
Station
Thal
Tria


BBC results: NBC at 320 kbit/s
mean and 95% confidence intervals
diffgrades: coded - reference
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
Cast
Clarinet
Eliot
Glock
Harp
Manc
Pipe
Station
Thal
Tria



6.6 Results from the NHK test site

6.6.1 Two-way ANOVA

A two-way fixed-effects ANOVA using the NHK data produced the following results:

Analysis of Variance for diffgrades: NHK Site, 16 listeners.

Source	Sum of squares	Degrees of freedom	Mean Square	F-ratio	P-level
Main Effects	25.065	12	2.089	2.708	0.001
CODEC	5.855	3	1.952	2.530	0.056
ITEM	19.210	9	2.134	2.767	0.004
2-Way Interactions	28.423	27	1.053	1.365	0.105
CODEC ITEM	28.423	27	1.053	1.365	0.105
Explained	53.488	39	1.371	1.778	0.003
Residual	462.803	600	0.771		
Total	516.291	639	0.808		

From the final column in this table it can be seen that, considering the data from all the codecs together, only the main effect Item was statistically significant for the NHK data. (However, differences between some codecs are revealed in Section 6.8.1).

The NHK data gives the following mean values for the overall performance of each codec. It must be noted, however, that item-to-item variations may also exist but these may not be apparent in these mean values.

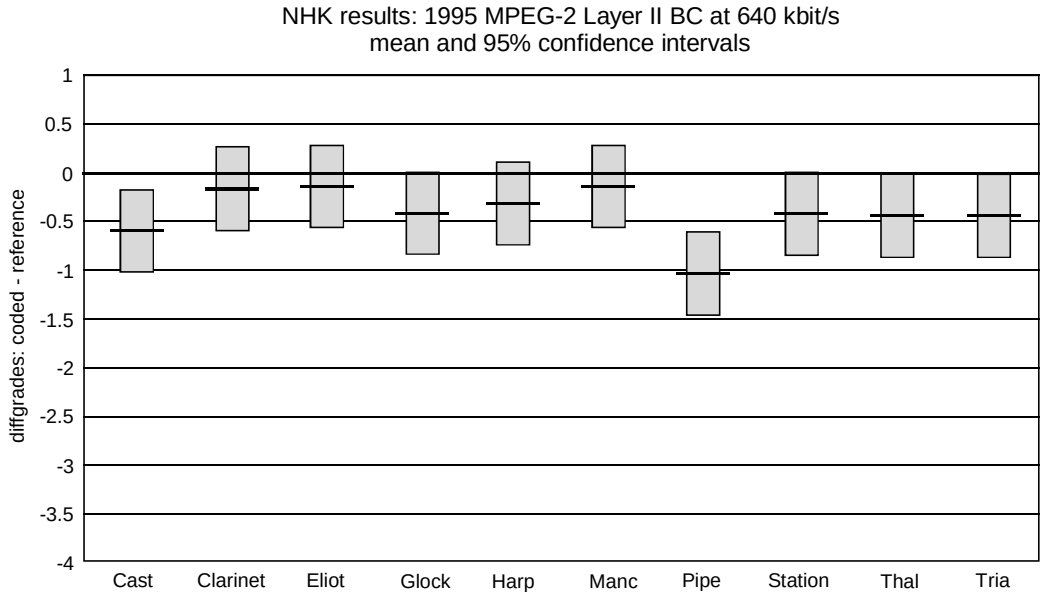
Mean diffgrades: NHK results

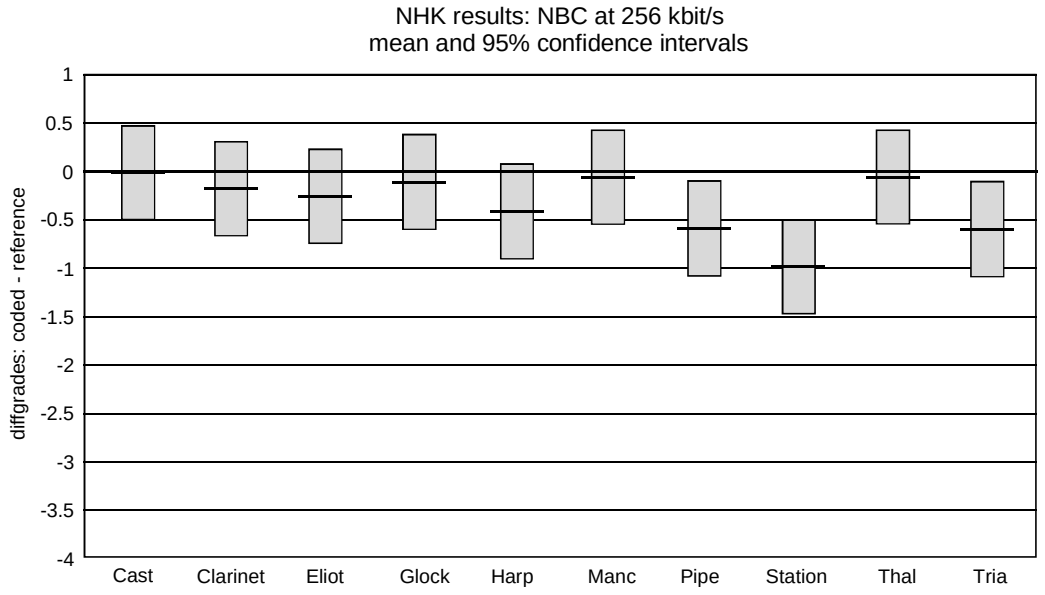
Codec	Samples	Mean diffgrade

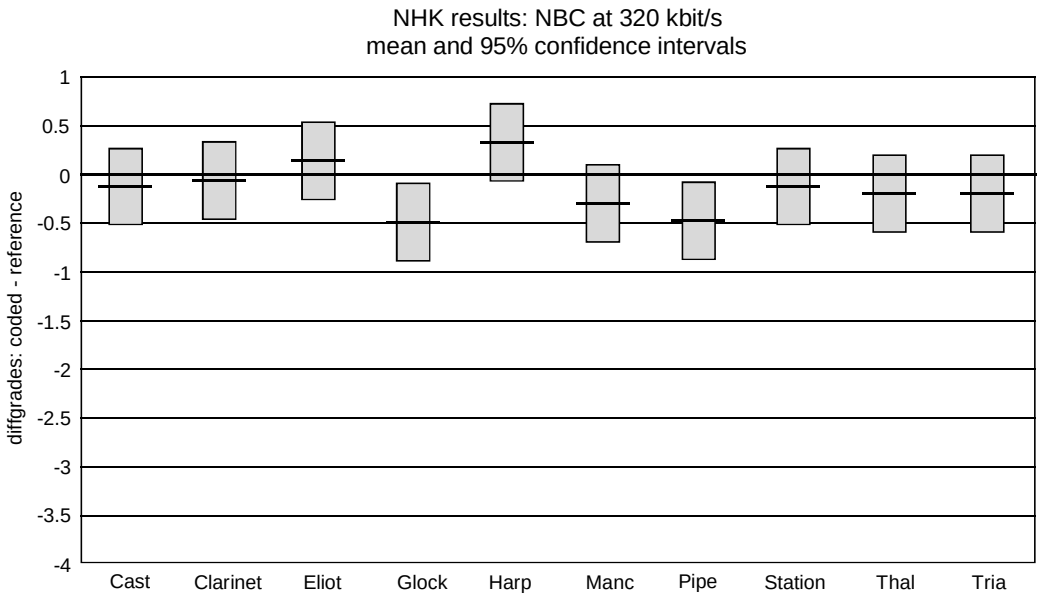
Layer II at 640 kbit/s	160	-0.41
NBC at 256 kbit/s	160	-0.33
NBC at 320 kbit/s	160	-0.15
NBC lc at 320 kbit/s	160	-0.26

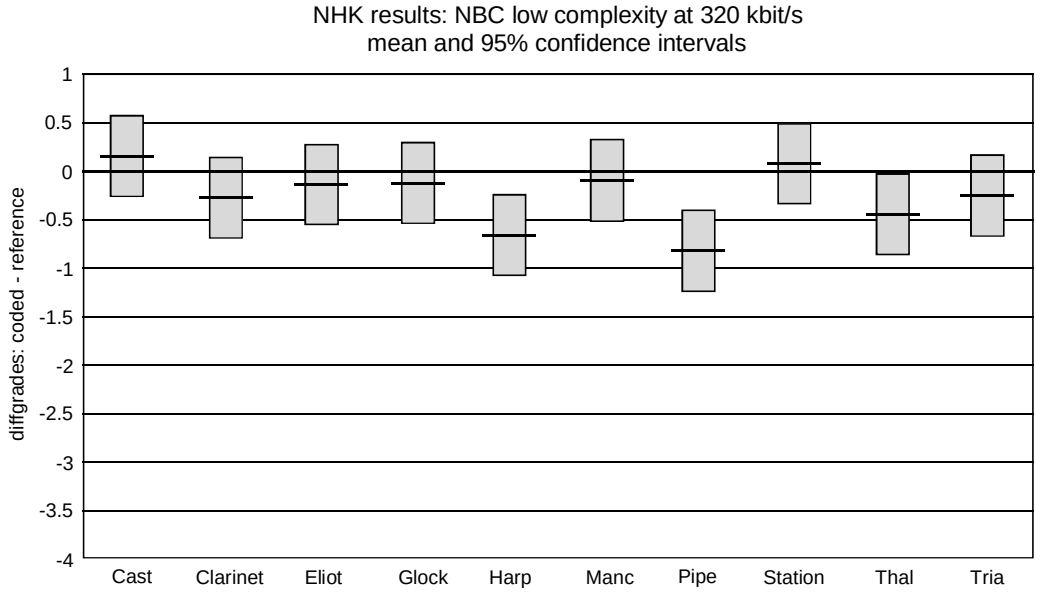
6.6.2 Estimates of means and 95% confidence intervals

The following four diagrams show the estimated means and two-sided 95% confidence intervals for each codec and excerpt. As the Levene's tests showed that the model assumptions were valid for all four codecs, the confidence intervals are computed from the one-way ANOVAs for each codec. Thus these confidence intervals possess equal length for each of the items. The results of the test on the model assumptions, the ANOVAs and the tabulated data for these diagrams are given in Annex G.









6.8 Comparisons of Codecs

6.8.1 MPEG-2 Layer II at 640 kbit/s and MPEG-2 NBC at 320 kbit/s

Question 4 of the test specification (see Section 6.1) asks if there are differences between the MPEG-2 Layer II BC codec at 640 kbit/s[4] and the MPEG-2 NBC codec at 320 kbit/s. In order to determine if significant differences exist, two ANOVAs (one for each site) were performed on the data for these two codecs. If these showed that the codec term is significant, either as a main effect or an interaction, then there are differences between the codecs. The results from these ANOVAs are:

[4] Note, the reader is reminded that these test results relate to the 1995 version of MPEG-2 Layer II BC and do not reflect any subsequent enhancements that may have occurred.

ANOVA: Comparison of Layer II and NBC (320) codecs: BBC results.

Source of Variation	Sum of Squares	DF	Mean Square	F - ratio	P - level
Main Effects	34.474	10	3.447	3.323	.000
CODEC	2.573	1	2.573	2.480	.116
ITEM	31.901	9	3.545	3.417	.000
2-Way Interactions	35.166	9	3.907	3.767	.000
CODEC ITEM	35.166	9	3.907	3.767	.000
Explained	69.640	19	3.665	3.533	.000
Residual	456.447	440	1.037		
Total	526.087	459	1.146		

It can be seen from the final column that the p-level for both the Item main effect and the Codec*Item interaction are below 0.05, i.e. they are statistically significant. For the BBC data, therefore, some items reveal significant differences between the two codecs.

For the NHK data, the ANOVA gives the following results:

ANOVA: Comparison of Layer II and NBC (320) codecs: NHK results.

Source of Variation	Sum of Squares	DF	Mean Square	F - ratio	P - level
Main Effects	20.074	10	2.007	2.897	0.002
CODEC	5.434	1	5.434	7.843	0.005
ITEM	14.640	9	1.627	2.348	0.014
2-Way Interactions	4.733	9	0.526	0.759	0.655
CODEC ITEM	4.733	9	0.526	0.759	0.655
Explained	24.807	19	1.306	1.884	0.015
Residual	207.862	300	0.693		
Total	232.669	319	0.729		

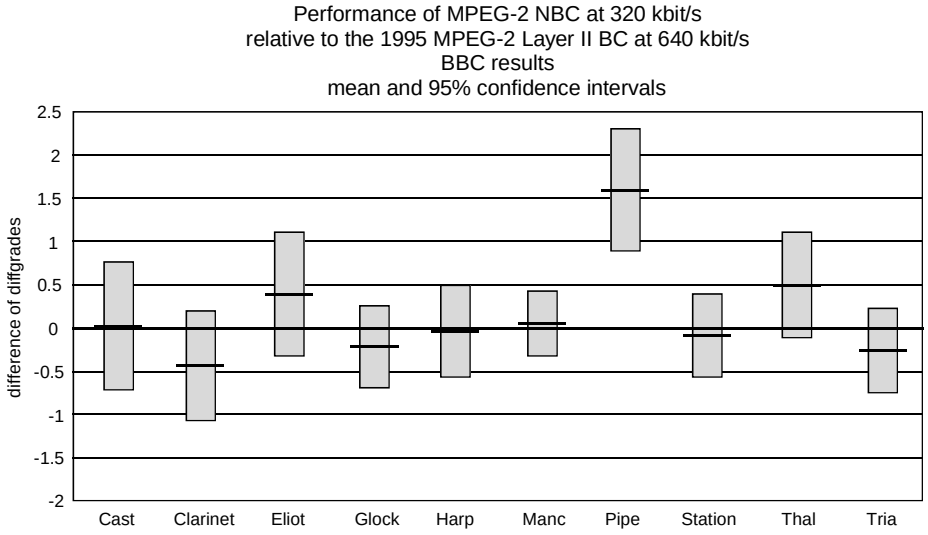
In a similar way, the above table reveals that, at the NHK site, both the Codec and the Item main effects are statistically significant.

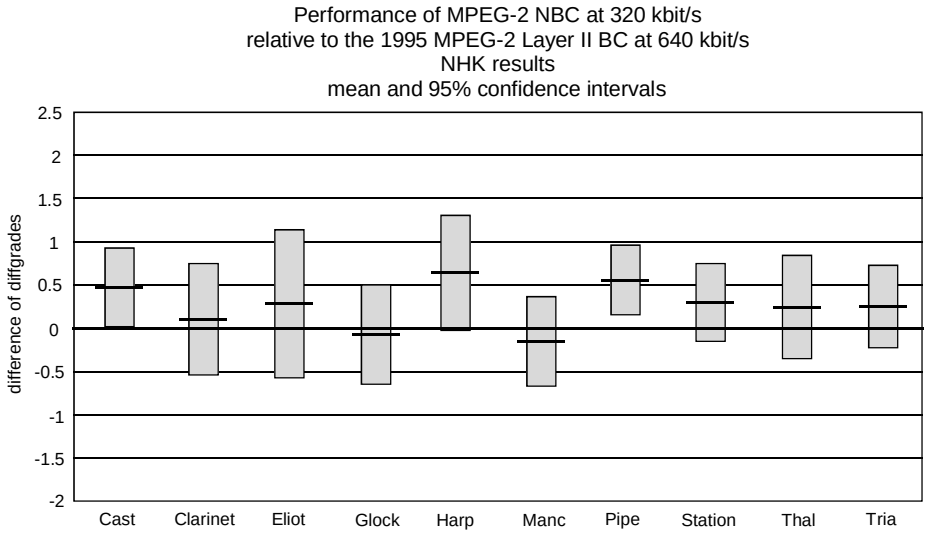
These results indicate that there are differences between these two codecs. This phenomenon can be seen in the diagrams in Sections 6.5.2 and 6.6.2 (and in the tables of means in Annex G), with the "NBC at 320 kbit/s" diffgrades generally out-performing (i.e. being closer to zero) the "Layer II at 640 kbit/s" diffgrades.

[Note the marked difference in residual mean squares in the above tables. This shows clearly the heterogeneity of variances between the BBC and NHK sites, and supports the approach of analysing the two sites separately.]

The differences of the diffgrades for these two codecs were calculated, item by item, and these are shown below for each test site. Note that in these diagrams, a positive value indicates that the MPEG-2 NBC codec was awarded a better diffgrade than the

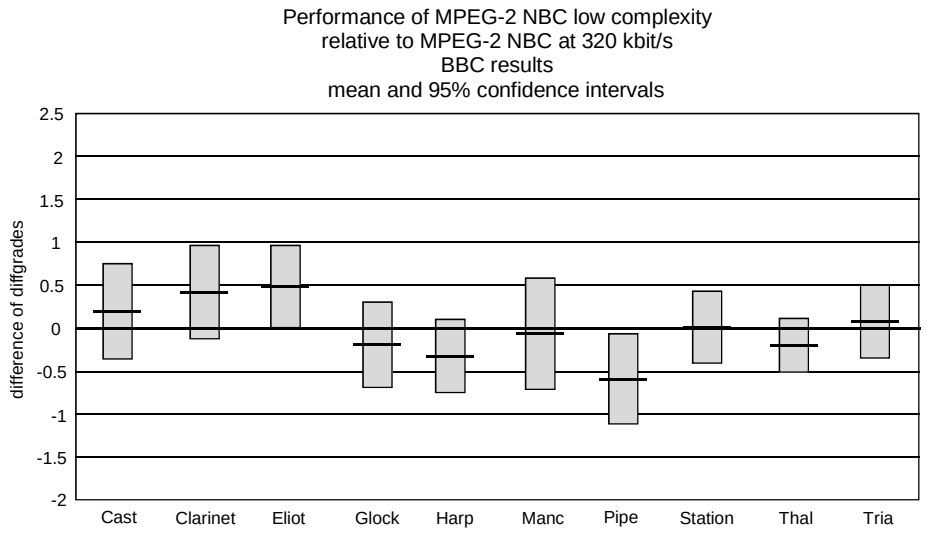
MPEG-2 Layer II BC codec and vice versa. The data for these diagrams is given in Annex G.

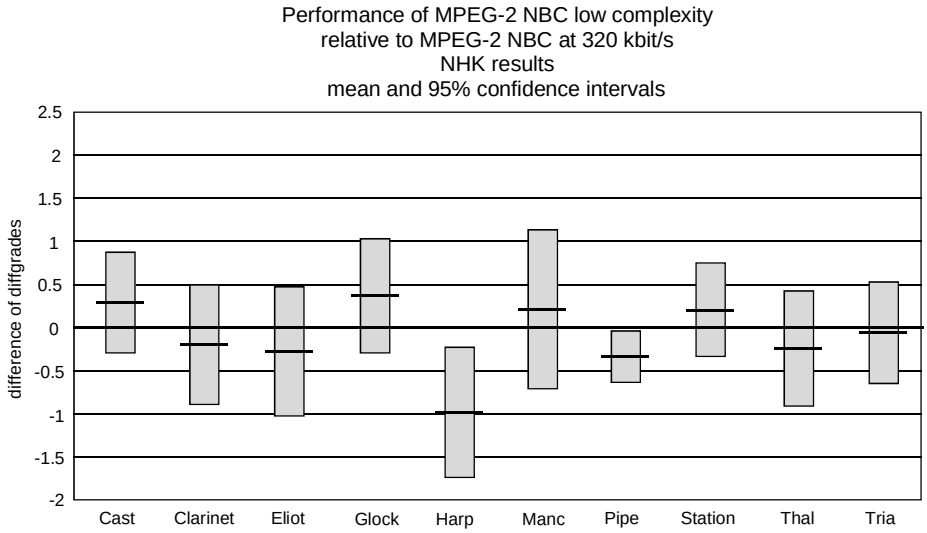




6.8.2 MPEG-2 NBC at 320 kbit/s and MPEG-2 NBC low complexity

As the performance of the MPEG-2 NBC low complexity implementation appears to be significantly better than suggested at the Tampere MPEG meeting, a comparison with MPEG-2 NBC at 320 kbit/s has been carried out. The differences of the diffgrades for these two codecs were calculated and these are shown below for each test site. Note that in these diagrams, a positive value indicates that the MPEG-2 NBC low complexity codec was awarded a better diffgrade than the MPEG-2 NBC codec at 320 kbit/s and vice versa. The data for these diagrams is given in Annex G.





6.9 Performance of MPEG-2 NBC at 320 kbit/s according to the EBU definition

Question 7 of the test specification (see Section 6.1) asks if the performance of the NBC codec at 320 kbit/s achieved 'indistinguishable quality' according to the EBU definition [21]. This section describes the analysis performed to answer this question.

For each site and each codec, one-way ANOVAs on items were performed separately for the test score and the reference score, yielding pooled error standard deviations. Thus, the lower end point of a 95% confidence interval for the mean reference score per item and the upper end point of a 95% confidence interval for the mean test score per item were calculated. This gave a cut-off value for the mean diffgrade per item. If the diffgrade for an item is less than the cut-off, then the confidence intervals do not overlap and the codec fails for that item according to the EBU criterion.

If the codec fails more than 3 items, it is deemed to have failed overall. If the codec fails 1,2 or 3 items, then it fails overall if the any of the ratios (upper end point of 95% confidence interval for item test) / (upper end point of 95% confidence interval for reference test) < 0.85.

The table below shows the results of these calculations for each test site.

Site / Codec	Cut-off	Items failing initially	Ratio (if appropriate)
Site: BBC			
Layer II at 640	-0.5434	Eliot	
		Harp	
		Pipe	
		Thal	
NBC at 256	-0.5651	Clarinet	
		Glock	
		Harp	
		Pipe	
		Station	
		Triangle	
NBC at 320	-0.4717	Clarinet	0.89
		Harp	0.90
		Triangle	0.93
NBC lc at 320	-0.4934	Harp	0.84
		Pipe	0.86
Site: NHK			
Layer II at 640	-0.5055	Cast	0.92
		Pipe	0.83
NBC at 256	-0.6093	Station	0.83
NBC at 320	-0.4912	Glock	0.91
NBC lc at 320	-0.5161	Harp	0.88
		Pipe	0.85

Thus, only the MPEG-2 NBC codec at 320 kbit/s passes the EBU criterion at both sites. The MPEG-2 NBC low complexity codec passes the EBU criterion at the NHK site and is borderline at the BBC site. The other codecs, MPEG-2 Layer II BC at 640 kbit/s and MPEG-2 NBC at 256 kbit/s, fail at both sites.

It should be pointed out that at neither site were there 40 or more subjects (as laid down by EBU).

The rationale for these EBU criteria is not clear.

6.10 Ranking of the codecs

To determine if a relative ranking of the codecs could be determined (question 8 in Section 6.1), the following two analyses were carried out.

Least significant difference analysis of codec means by site

Using the overall error mean squares from the analyses of variance, the smallest difference in codec mean that is statistically significant at the 95% confidence level is 0.1921 for the BBC site and 0.1925 for the NHK site.

Thus, for the BBC site codecs "NBC at 320", "NBC low complexity at 320" and "Layer II at 640" form one group (better) and codec "NBC at 256" is significantly different from the others.

For the NHK site, codec "NBC at 320" is not significantly different from codecs "NBC low complexity at 320" and "NBC at 256", but is significantly different from codec

"Layer II at 640". Codecs "NBC low complexity at 320" and "NBC at 256" are not significantly different from any other codec.

So this does not give a particularly clear picture.

A simple method of comparison

The following shows the number of items for each codec at each site for which the 95% confidence interval for mean diffgrade (a) contained 0 and (b) contained -1 or less.

Site / Codec	Number with 0 in Conf. Int.	Number with -1 or less in Conf. Int.
Site : BBC		
Layer II at 640	6	3
NBC at 256	2	8
NBC at 320	6	1
NBC lc at 320	7	2
Site : NHK		
Layer II at 640	5	2
NBC at 256	7	3
NBC at 320	8	0
NBC lc at 320	7	2

This, perhaps, indicates a rough ordering of codecs: "NBC at 320" and "NBC low complexity at 320" (best), followed by codec "Layer II at 640" and finally codec "NBC at 256" for the BBC site.

For the NHK site this ordering is: codec "NBC at 320" (best), followed by codec "NBC low complexity at 320", then codecs "Layer II at 640" and "NBC at 256".

Interestingly, these results agree quite well with the EBU criteria presented in Section 6.9.

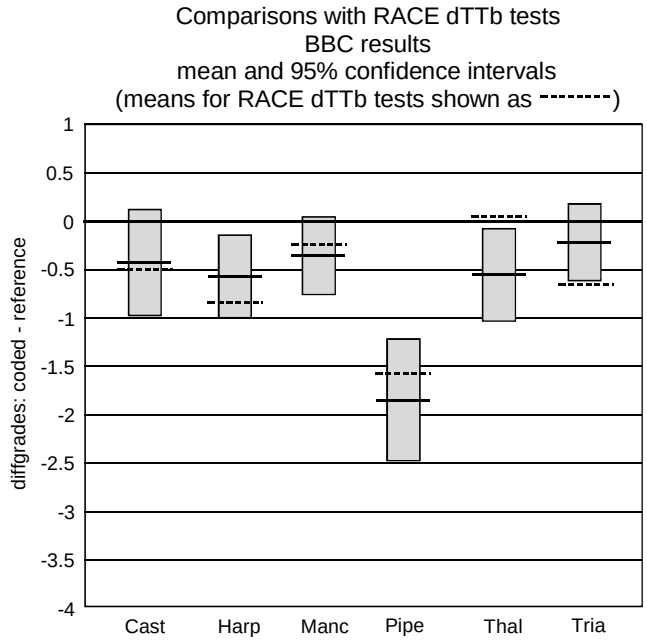
6.11 Tests on model assumptions

The Shapiro-Wilk test was applied to assess the normality of the initial data and this showed that statistical significance was reached for all codecs (see Annex G). So, strictly speaking, the underlying normality assumption for the ANOVAs has been violated. However, ANOVA is robust to modest departures from normality, and inspection of the raw data indicated that it was not grossly non-normal. The relatively small p-values are what one would expect, given the relatively large amount of data and the power of the test to detect even small departures from normality for relatively large data sets.

7. Comments on test results.

7.1 Comparison with earlier tests

The BBC results for the low anchor presentations (Section 6.3) show consistency between these tests and the MPEG '94 tests [1]. For a direct comparison between these tests and the RACE dTTb tests [4], the results for the 1995 MPEG-2 Layer II BC codec at 640 kbit/s are shown in the diagram below for both tests.



These show good agreement with four of the six mean values from the dTTb tests lying within the confidence intervals for these tests. For one of the six which lies outside these confidence intervals, namely triangle, the converse applies, i.e. the mean diffgrade from this test lies within the 95% confidence interval of the dTTb results.

The apparent differences between the means for the Thalheim and Triangle items were further investigated by applying a two-sided t-test. This indicated that, at the 95% confidence level, there is no significant difference between the means from each test (p-levels: 0.11 and 0.20 respectively).

7.2 Summary of answers to initial questions

The answers to the questions posed in the test specification [16] (see Section 6.1) can be summarised as follows:

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

   An assessment of listener reliability has been performed and only reliable data has been used in the analysis (see Section 6.2).

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

   Yes. Furthermore, performance below the level of transparency could be detected and these tests also appear to be reasonably consistent with earlier tests.

3) Is there any distinction between the two test sites?

   Yes; a three-way ANOVA was performed to check this and differences were found (see Section 6.4).

4) Is the performance of MPEG-2 NBC at the default bitrate *[320 kbit/s]* equal to or better than the performance of *[the 1995 version of]* MPEG-2 BC Layer II at 640 kbit/s?

   Differences between these codecs were revealed. Generally, the performance of the MPEG-2 NBC codec at 320 kbit/s appears to be better.

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

   Sections 6.5.2 and 6.6.2 of this report show the performance of each codec for each of the programme items.

6) Is the performance of the coding of NBC at the default bitrate *[320 kbit/s]* distinguishable from the original signal?

   The diagrams shown in Sections 6.5.2 and 6.6.2 indicate that the MPEG-2 NBC codec at 320 kbit/s is statistically distinguishable from the original signal for some excerpts.

7) Is the performance of NBC at the default bitrate *[320 kbit/s]* achieving 'indistinguishable quality' in the EBU definition [21] of that phrase?

   Yes, at both test sites. (However, in each case, fewer than the recommended 40 listeners participated). See Section 6.9.

8) What is the relative ranking of the codecs tested?

   A clear ranking of the codecs is difficult to determine as their grouping differs between the test sites. However, generally, MPEG-2 NBC at 320 kbit/s and MPEG-2 NBC low complexity performed better than the 1995 version of MPEG-2 Layer II BC at 640 kbit/s and MPEG-2 NBC at 256 kbit/s; see Section 6.10 for further details.

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

   A comparison of MPEG-2 NBC at 320 kbit/s and MPEG-2 NBC low complexity has been included, see Section 6.8.2.

7.3 Further observations on the tests and the results

- The results for all the codecs show very good performance. During the tests, most subjects found it necessary to listen to each trial many times because of the difficulty in identifying the coded version.
- All the variants of MPEG-2 NBC coding which were tested, achieved approximately the same 5-channel performance or better at half the bitrate of the 1995 version of the MPEG-2 Layer II BC codec.
- At both test sites, the MPEG-2 NBC codec at 320 kbit/s achieved diffgrades better than -0.7 for all of the test excerpts (i.e. better than grade 4.3 on the impairment scale). Only two of these excerpts, clarinet and harpsichord at the BBC site, gave mean diffgrades worse than -0.5 and only one excerpt, clarinet, gave rise to a 95% confidence interval which crossed below the diffgrade value of -1.0.

- The MPEG-2 NBC low complexity implementation at 320 kbit/s achieved mean diffgrades at both test sites better than -1.0, i.e. better than grade 4.0, on the impairment scale.
- With the implementations tested, MPEG-2 NBC low complexity at 320 kbit/s is only marginally worse than MPEG-2 NBC at 320 kbit/s.
- Care should be exercised when comparing the performance of the different implementations of the MPEG-2 NBC codecs as they had different features enabled in addition to the differing bitrate or level of complexity.
- Where test stimuli, low anchors or MPEG-2 Layer II BC, had been previously assessed in earlier tests of this nature, the results from this series of tests show very similar results to those previously published.
- Further evaluations of the MPEG-2 NBC coders may be warranted once further coding optimisation has been carried out.
- No assessments have yet been reported on the two-channel stereo performance of the MPEG-2 NBC codecs. If two-channel reproduction is to be achieved by simulcasting using an existing stereo coder, then results from stereo coding tests can be assumed to be relevant, but the bitrate will increase accordingly. If the stereo version is to be created by downmixing of the 5-channels delivered by MPEG-2 NBC coding, then the bit rate will be as reported here, but subjective assessments of the stereo performance should be made.

8. References

1. Feige, F. & Kirby, D.G., 1994, “Report on the MPEG/Audio multichannel formal subjective listening tests”. MPEG document ISO/IEC JTC1/SC29/WG11/N0685. March 1994.
2. Meares, D. & Kirby, D., 1995. “Report on brief assessments of the performance of ISO/MPEG-2 Layer 3 at 7*128kb/s. ISO/IEC JTC1/SC29/WG11/N0895”. March 1995.
3. Kirby, D., 1995. “Brief assessments of the performance of an ISO/MPEG-2 Layer 2 audio codec operating at 896 kb/s”. ISO/IEC JTC1/SC29/WG11/N0974. July 1995.
4. Feige, F. & Kirby, D., 1996. “MPEG-2 Backwards compatible codecs Layer II and Layer III: RACE dTTb listening test report”. ISO/IEC JTC1/SC29/WG11/N1229. March 1996.
5. Audio Subgroup, 1994. “Requirements for Preliminary Proposal and Draft Requirements for Detailed Technical Proposal for the MPEG-2 Non-backward Compatible (NBC) Extension”. ISO/IEC JTC1/SC29/WG11/N0746. July 1994.
6. Schreiner, P.,1995. “Blockwise analysis of the NBC codecs according to RM 0”. ISO/IEC JTC1/SC29/WG11/MPEG95-0035. March 1995.
7. Audio Subgroup, 1995. “Definitions and interface description for the time to frequency mapping of RM1”. ISO/IEC JTC1/SC29/WG11/N0933. March 1995.
8. Audio Subgroup, 1995. “Definitions and interface descriptions for RM2”. ISO/IEC JTC1/SC29/WG11/N0982. July 1995.

9. Dietz, M., 1995. “Definitions and interface descriptions for NBC RM2”. ISO/IEC JTC1/SC29/WG11/N1095. November 1995.

10. Bosi, M., 1996. “MPEG-2 Audio NBC (13818-7) Reference Model 3 (RM3)”. ISO/IEC JTC1/SC29/WG11/N1132. January 1996.

11. Bosi, M., et al, 1996. “MPEG-2 Audio NBC (13818-7) Working Draft”, Version. ISO/IEC JTC1/SC29/WG11/N1200. March 1996

12. Meares, D. & Kim, S-W., 1995. “NBC time/frequency module subjective tests: overall results”. ISO/IEC JTC1/SC29/WG11/N0973. July 1995.

13. Meares, D. & Kim, S-W., 1996. “NBC Reference Model 2 subjective tests: overall results”. ISO/IEC JTC1/SC29/WG11/N1135. January 1996.

14. Lueck, C. & Ali, M., 1996. “Preliminary NBC Reference Model 3 subjective tests: overall results”. ISO/IEC JTC1/SC29/WG11/N1212. March 1996.

15. Lueck, C., Ali, M., Thom, D., Meares, D., Schreiner, P. 1996. “NBC Reference Model 3 monophonic subjective tests: overall results”. ISO/IEC JTC1/SC29/WG11/N1279. July 1996

16. Feige, F., Meares, D. J., “Specification of MPEG-2 Audio NBC Formal Tests”. MPEG document ISO/IEC JTC1/SC29/WG11/N1281. July 1996.

17. ITU-R Recommendation BS.1116 “Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems”. Geneva (1994)

18. Chiariglione, L., 1996. “Report of the 35th meeting in Tampere, Finland.” ISO/IEC JTC1/SC29/WG11/N1270. July 1996.

19. Lueck, C., et al, 1996. “NBC Reference Model 4 stereophonic and multichannel subjective tests: overall results”. ISO/IEC JTC1/SC29/WG11/N1280. July 1996

20. Fielder, L., 1996. “Call for multichannel audio test sequences”. ISO/IEC JTC1/SC29/WG11/N1203. March 1996.

21. EBU, 1991. “Basic audio quality requirements for digital audio bit-rate reduction systems for broadcast emission and primary distribution”. CCIR document number TG 10-2/3, 28 October 1991. (Reproduced for convenience in Annex H.)

Annex A. Report of the Selection Panel.

Report of the Selection Panel for the ISO/MPEG NBC Listening Tests

Tasks assigned to the selection panel

1. Select the 10 most critical items by listening to a range of new and old programme material through all codecs. Repetition of similar material should be avoided, e.g. bell and triangle should not both be included. Also say which test items could be omitted if the test timetable so dictates.

2. Select 4 items for use in the training sessions. These should provoke the full range of artefacts expected in the tests and fairly represent the artefacts produced by each codec. The training items should preferably, but not necessarily, be a subset of the 10 test items.

   Say whether the proposed lower bit-rate versions are suitable for training, i.e. that they demonstrate clear artefacts.

3. Ensure that the selected codec/item combinations include a number which will be suitable as “low anchors”, i.e. which will produce average grades of approximately 3.5 but not less than 3.0.

4. Identify any codec/bit-rate combinations which should be excluded on the grounds of consistently poor quality, i.e. which will produce average grades below 2.5 (or possibly even 3.0).

5. Offer advice concerning the tests.

Conclusions

1. Selection of the 10 most critical items

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

No.	Name	Description
1	pitch_pipe	Pitch Pipe
2	harpsichord	Harpsichord
3	triangle	Triangle
4	cast_pan1	Castanets panned across the front, noise in surround
5	elliot1	Female and male speech in a restaurant, chamber music
6	mancini	Orchestra - strings, cymbals, drums, horns
7	station_master1	Male voice with steam train
8	clarinet_theatre	Clarinet in centre front, theatre foyer ambience, rain on windows in surround
9	thalheim1	Piano front left, sax in front right, female voice in centre
10	glock	Glockenspiel and timpani

“Glock” followed by “thalheim1” could be omitted if necessary.

Artefacts observed with the 10 selected critical items

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

No.	Piece	Artefacts
1	pitch_pipe	Distortion, Quantisation defects, Periodic modulation, Image quality, Noise
2	harpsichord	Distortion, Temporal distortion, Periodic modulation
3	triangle	Temporal distortion, Image quality, Non-periodic modulation, Periodic modulation
4	cast_pan1	Temporal distortion, Image quality, Non-periodic modulation, Quantisation defects
5	elliot1	Quantisation defects, Noise, Excess of high frequency
6	mancini	Quantisation defects, Non-periodic modulation, Image quality
7	station_master1	Quantisation defects, Periodic modulation, Distortion
8	clarinet_theatre	Distortion, Quantisation defects, Periodic modulation
9	thalheim1	Quantisation defects, Distortion, Image quality.
10	glock	Distortion, Periodic modulation, Temporal distortion

Artefact categories for each codec

This table contains a list of the main artefacts found in each codec (at the “high” bit-rate) for each item. The artefacts are listed in approximate order of severity. Those in bold are major artefacts while those in italics are minor artefacts. See the Appendix for the numbers corresponding to the artefact categories.

Item/Codec	A	B	C	D
pitch_pipe	**8**, 1, 4, 11	8, 12	8, 4, 1, 11	**8**, 1, 4, 11
harpsichord	5, 11	8, *12*	4, 3, 11	1, 8
triangle	11, *1*, *4*	8, *11*	8, 4	8, 5, 4
cast_pan1	**7**, 1, 11	*4, 12*	8, 4, *3*	8, 4, 5
elliot1	**1**, *11, 12*	1, *4*	1, 3	1, 8, *11*
mancini	**1**, 8, 11	*8, 12*	1, *3*	**1**, 8
station_master1	**1**, 11	8, *1*	1, 4	1, 8
clarinet_theatre	1, 8	8	1, 3, 8	8
thalheim1	8, 1, 11	1, 8, *12*	3, 8, *1, 4*	1, 8
glock	7, 5, 1, *11*	5, *7*	5, 4, *3*	7, 5

Summary of main characteristics

Codec A was characterised by Quantisation Defects, Image Quality problems, Distortion and Temporal Distortion.

Codec B was characterised mainly by Distortion with the addition of some Noise.

Codec C was characterised by Quantisation Defects and Distortion, with an Excess of High Frequency and Periodic Modulation.

Codec D was characterised by Distortion and Quantisation Defects.

2. Training Items

The following four of the selected ten most critical items are recommended for training of the test subjects. These items were found to represent almost all of the impairments detected in the codecs. Although “thalheim1” may be dropped from the test if the requirement is for only 8 items, it contains both speech and individual instruments and is therefore very useful as a training item.

No.	Name
1	harpsichord
2	triangle
3	mancini
4	thalheim1

3. Low anchors

After an evaluation of all codecs at “high” bit-rate the selection panel did not find any items that were suitable as a low quality anchor in the test. At “low” bit-rate (which was intended for training) a few codec/item combinations were found to be suitable as low anchors.

In addition, one of the codecs was evaluated at an even lower bit-rate with the most critical items. Three of the items were found to be too poor in quality to use as low anchors in the test. Hence the selection panel recommends that this even lower bit-rate codec not be included in the test.

4. Poor quality codecs

None of the codecs at “high” or “low” bit-rate should be rejected on grounds of consistently poor quality.

5. Advice concerning the test

The selection panel noted that it would have been helpful for time to have been allocated for preselection from the large number of test items offered.

The selection panel further noted that although they were asked to ensure that the test items included “low anchors”, they could not fulfil this requirement with the codecs provided. An independent assessment of the codecs prior to the selection process might have been helpful in this context.

APPENDIX

Details of the Selection Process

Contents

1. Listening room and technical equipment

Listening took place in Listening Room 2 at BBC Research & Development, Kingswood Warren. This is the same room in which the actual tests will take place. The room is acoustically treated and is approximately 4.4 m x 5.5 m and 3.0 m high. It is equipped with five Rogers LS5/8 loudspeakers.

Test material was replayed from a Tascam DA-88 digital audio tape recorder via a Yamaha DMC1000 desk and Prism Dream DA-1 digital-to-analogue converters, all located in an adjacent room.

Software running on a Unix workstation provided control of the tape recorder (using RS422). A list of items on the tape was displayed on a console in the listening room and, by positioning a cursor, the tape could be cued to any item, played, stopped etc. The item order on tape was: Reference, A, B, Reference, C, D, Reference.

Tapes were prepared using a Sonic Solutions disc-based audio editor. Test items were loaded from a Unix workstation onto the Sonic (in AIFF format) and locally written software was used to generate the required edit lists. The edit lists were then replayed and recorded onto Tascam DA-88. In this way, tapes were prepared according to the requirements of the selection panel as the selection process proceeded.

2. Item list reduction process

The reduction process started with some preliminary listening at the BBC, just before the selection panel met. The aim was to remove items from consideration that were obviously not critical. Two versions of the codecs, at different bit-rates, were available: “low” and “high”, the former for training and the latter for the tests. The items were listened to at the “low” bit-rate for all codecs. Items marked “not (very) critical” were not considered further. 53 of the 94 available items were auditioned in this preliminary listening.

The codecs were found to be of quite high quality and so the suggestion was made that the selection panel initially listen to all of the critical items from the previous dTTb test. After listening to each item the panel discussed all of the artefacts observed, repeating the item if necessary. The panel marked each item for overall criticality and suitability to serve as a low anchor. Items marked “not (very) critical” were not considered further.

The panel listened, at the “low” bit-rate, to the dTTb critical items first, then the survivors of the preliminary listening followed by the remaining items; non-critical items were then dropped (1st step). The surviving items were then listened to at the “high” bit-rate; again non-critical items were dropped (2nd step).

The next phase marked those as most critical (3rd step). The final list of ten consists of the most critical items, with consideration given to balancing the content of the items (4th step). This process is tabulated in the "List for Selection of Test Excerpts" (Section 4 of this Appendix).

3. Impairment Categories Table

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

No.	Artefact Category	Explanation
1	Quantisation Defects	defects associated with insufficient resolution, e.g. granular distortion
2	Loss of High Frequency	lack of high frequencies
3	Excess of High Frequency	excess of high frequencies or associated effects, e.g. sibilance or hissing
4	Periodic Modulation Effects	periodic variations such as warbling, pumping, or twitter
5	Non-periodic Modulation Effects	effects associated with transients, e.g. splats or bursts
6	Level Change	change in level of a source effect (such as applause)
7	Temporal Distortion	pre- and post-echoes, smearing
8	Distortion	harmonic or inharmonic distortion
9	Extra Sounds	spurious sounds not related to the material
10	Correlation Effects	crosstalk between channels, e.g. bleeding or inter-channel correlation
11	Image Quality	all aspects including spreading, movement, stability and phase related effects
12	Noise	increased noise of uniform nature, e.g. background noise

4. List for Selection of Test Excerpts

Name	1st Step	2nd Step	3rd Step	4th Step
AES_Berlin	x			
C_sabre	x	x		
applause				
approaching_tunnel				
beethoven				
bell	x	x	x	
carneval				
castpan1	x	x	x	x
castpan2	x	x	x	
chostakovitch				
circus1				
clarinet+effects				
clarinet_theatre	x	x	x	x
doors+whistle	x	x	x	
drinks_inside	x	x		
driver				
drums				
drums1	x	x		
drums2				
elliot1	x	x	x	x
elliot2				
flute_in_rev_room				
flute_pan	x	x		
fountain_music	x			
franck1				
franck2				
genzmer2	x	x	x	
glock	x	x	x	x
guitar				
harpsichord	x	x	x	x
harpsiglock_melody	x	x		
harpsiglock_scale	x	x		
harpsiwood	x			
indie2	x			
infinito1	x	x		
interview	x	x		
jackson1				
jackson2	x	x		
jazz1	x			
jazz2	x	x		
jazz_finale	x			
klein1	x	x		
klein2				
klein3				
ligetti1	x			
ligetti2	x			
mancini	x	x	x	x
mendelssohn1	x	x		

Name	1st Step	2nd Step	3rd Step	4th Step
mendelssohn2[5]	x	x		
mussorgski	x	x		
oboe_pan	x			
orchestra				
organ	x			
party_talk	x	x		
piano	x			
pitch_pipe	x	x	x	x
pops				
ravel1	x			
ravel2	x			
rock_concert	x	x		
rock_fiddle	x	x		
rock_fiddle_v2				
roussel1	x			
roussel2	x			
sax	x	x		
saxo	x	x		
seawash	x	x		
station_master1	x	x	x	x
station_master2	x			
takacs1				
takacs2				
takacs3				
takemitsu	x			
tattoo1	x			
tattoo2				
tattoo3				
tennis				
thalheim1	x	x	x	x
thalheim2	x	x		
thalheim3				
thalheim4	x	x		
thalheim5				
tipsy	x	x		
tower1				
tower2				
train2	x			
train_passing	x			
train_under_bridge	x	x		
triangle	x	x	x	x
vilia	x			
violin1				
violin2				
violin3				

[5] A member of the selection panel noted that this item was, in fact, by Mussorgsky.

Annex B. Instructions to listeners.

MPEG-2 subjective tests on multichannel audio systems

Guidance information for subjects.

(Listeners are asked to read this in advance of their individual test period!)

General

The audio group of the ISO/MPEG-2 project is working on various techniques to improve the quality of multichannel audio coding systems. This work has now reached the stage where it is appropriate to assess the audio quality which can now be achieved by these latest versions of MPEG-2 multichannel coding under various operating conditions.

The systems under test are all five-channel 'surround-sound' systems with 3 front channels Left, Centre, Right (L, C, R), and 2 surround channels Left-surround, Right-surround (LS and RS) (towards the rear).

The multichannel performance of the systems will be assessed at two test centres: the BBC's R&D Department at Kingswood Warren, UK and at NHK, Science and Technical Research Labs, Japan. In both cases the tests will be carried out without pictures. There will be guidance from the test centres for the listeners throughout the test period.

2. Test procedure

In order to ensure consistent results, the test procedure for these tests follows the ITU-R Recommendation "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems".

The tests have two phases; firstly, a training phase involving groups of test subjects and then the grading phase where subjects conduct the tests individually.

2.1 Training phase.

The purpose of the training phase is to allow listeners to identify and become familiar with potential distortions and artefacts produced by the systems under test. You will also become familiar with the test procedure. This training phase is carried out in groups of 3 or 4 listeners and during this time you can comment on the items and discuss the artefacts heard with each other. After this training, you should know "what to listen for".

You have up to three hours for this training phase, which will also include an opportunity to hear the test items which will be used in the test phase.

Although an exchange of views on what you hear is expected during this training session, it is important that you should not discuss with the other listeners the grade that you, as an individual, would award, as this is a personal interpretation of the severity of the artefacts heard.

2.2 Test phase

The test phase will be carried out individually in test sessions each lasting about 25 to 30 minutes. In each trial, you will hear three versions, labelled “Ref”, “A” and “B” on the computer screen. “Ref” is always the reference (original) version against which both the “A” and “B” versions are to be compared and graded. One of “A” and “B” is a processed (coded/decoded) version and the other is a hidden reference (identical to the “Ref” version).

You are not told which of “A” and “B” is the processed version and which is the hidden reference, and this will change randomly from one trial to the next. You can switch freely between “Ref”, “A” or “B” at any time. This should allow a detailed comparison between “Ref”, “A”, and “B”. The audio excerpts can be played repeatedly until you are confident about your decision.

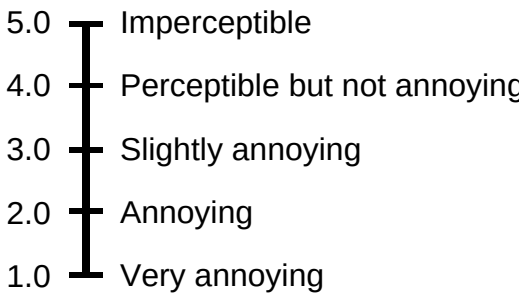
You are asked to judge the “Basic Audio Quality” of the “A” and “B” versions in each trial. This attribute is related to any and all differences between the reference and the coded/decoded programme excerpt. Note: Any difference between the reference and the coded/decoded programme excerpt is to be considered as an impairment.

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

It includes such things as harmonic distortions, added ‘pops’ or ‘cracks’, quantisation noise in subbands, pre-echoes (or other time smearing effects), changes in loudness, changes in timbre, changes in spatial presentation, changes in background noise or reverberance. Anything else that the listener detects as a difference must be included in his/her overall rating.

In each trial, you are asked to rate the perceived difference (if any) between

“Ref” and “A” and also the difference between “Ref” and “B” using the grading scale:



Two grades must be given on each trial, one for “A” and one for “B”. At least one grade of “5” must be given each trial since one of “A” or “B” is the hidden reference.

Please input your grades on the computer at the end of each trial, preferably with one decimal place.

It should be noted that the order of presentation of the test blocks and the position of the hidden reference is randomised for each test subject. Comments made by one listener during the test phase will therefore not be relevant to the perceptions of other listeners.

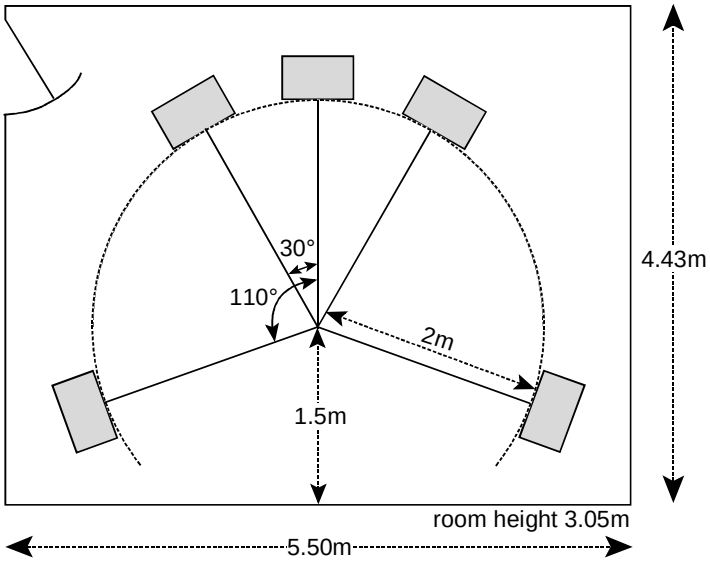
3. Time schedule for the tests.

Four listeners are scheduled for each test period of two days. Each of these test periods will start in the morning of the first day with the main training session, for all four of these subjects together. The tests start in the afternoon of the first day with each subject grading at least one test block. The remaining test blocks are to be completed on the second day, although the exact schedule for the day can be decided between the listeners beforehand, to give some flexibility in individual start and finish times.

Annex C. BBC test site: Listening Room Conditions and Equipment

Listening Room Conditions

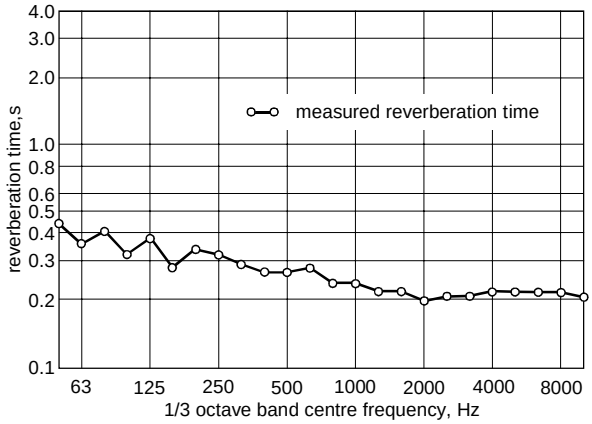
Room Dimensions



- Length: 4.43m
- Width: 5.50m
- Height: 3.05m
- Floor area: 24.40m²
- Volume: 74.30m³

Listening Room 2, BBC R&D Department, Kingswood Warren

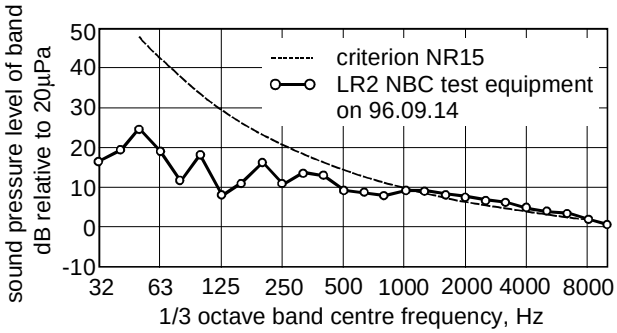
Reverberation time



BBC Listening Room 2

The mean reverberation time between 200 Hz and 2 kHz is 0.27s, which is in accordance with the ITU-R recommendations for this size of room.

Background noise



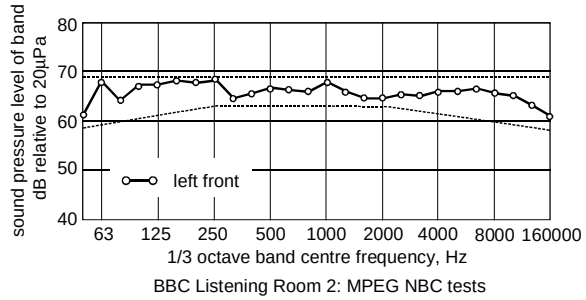
BBC Listening Room 2: MPEG NBC tests

The noise level at the reference listening position meets the noise criterion NR 16. (The high frequency noise just above the NR15 criterion is attributable to the audio power amplifiers.)

Frequency response measurements

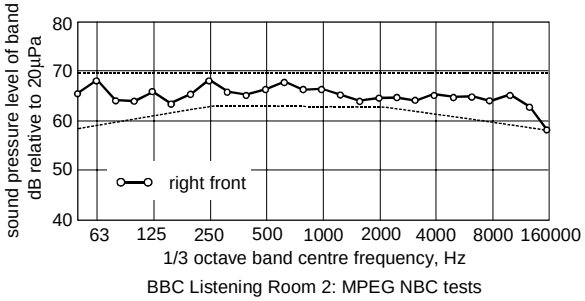
Measured at the reference listening position in 1/3 octave frequency bands using pink noise.

Left Front



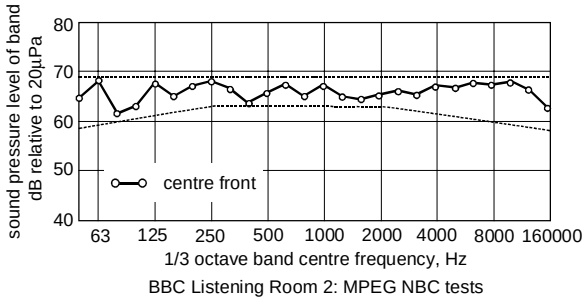
BBC Listening Room 2: MPEG NBC tests

Right Front



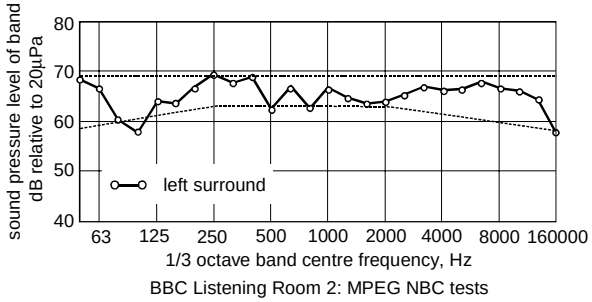
BBC Listening Room 2: MPEG NBC tests

Centre Front



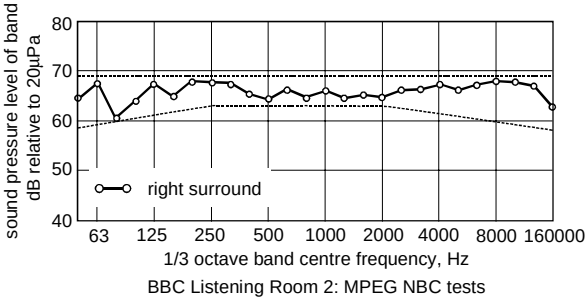
BBC Listening Room 2: MPEG NBC tests

Left Surround



BBC Listening Room 2: MPEG NBC tests

Right Surround

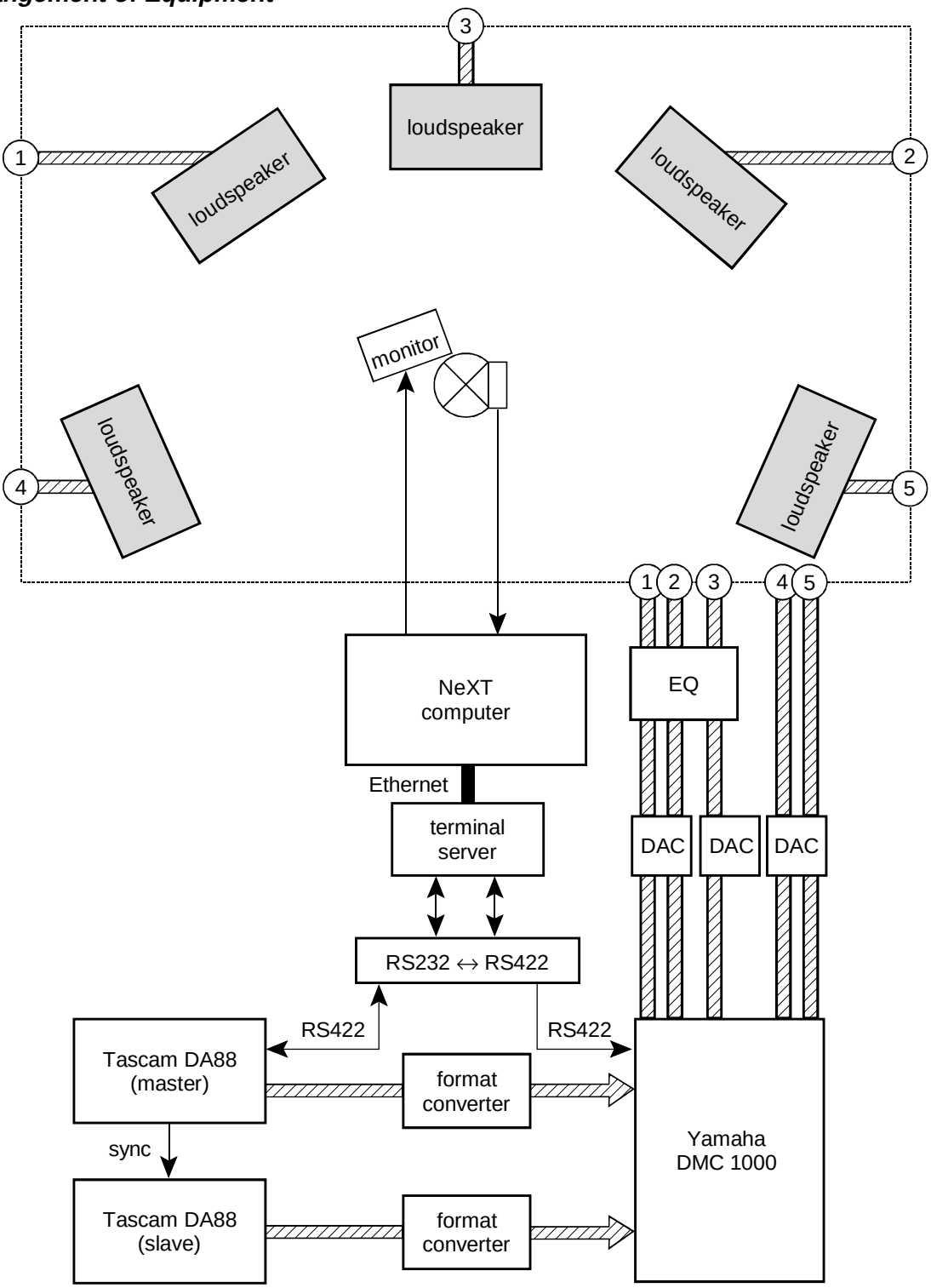


BBC Listening Room 2: MPEG NBC tests

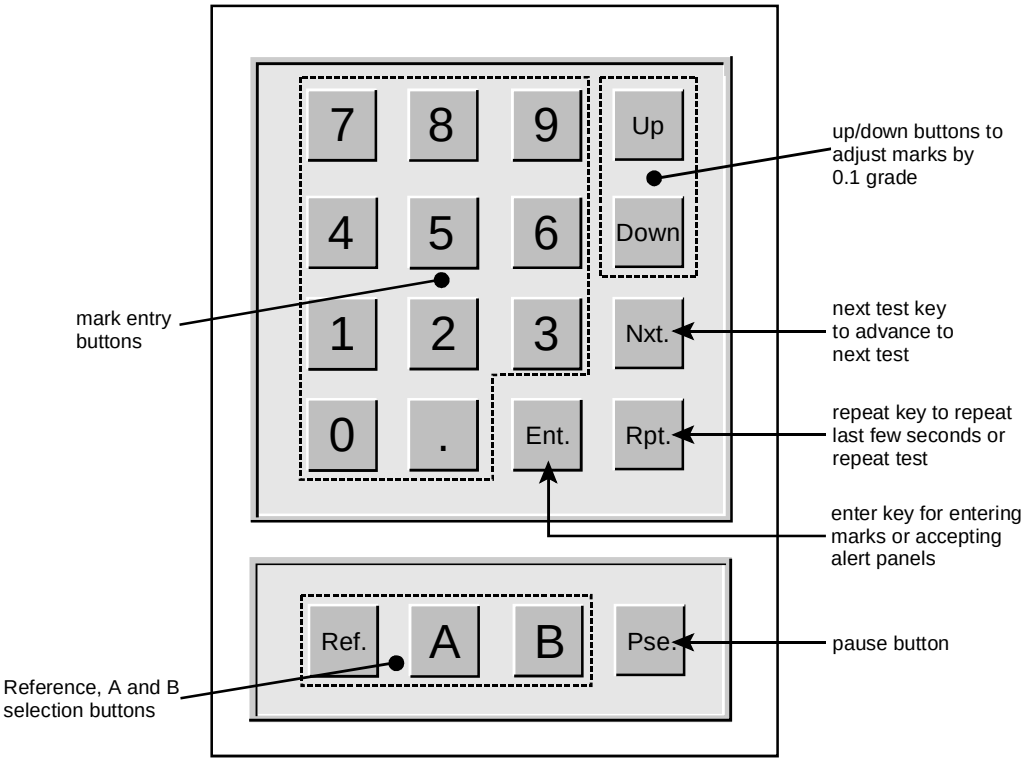
Equipment list.

	Device	Manufacturer	Description
2	DA88	Tascam	Digital audio recorders
1		Custom	Tascam to Yamaha digital audio format converter
1	IF88-AE	Tascam	Tascam to AES/EBU digital audio format converter
1	DMC1000	Yamaha	Digital mixing console
3	DA-1	Prism	Two-channel DAC
3			1/3 octave graphic equalisers
5	LS5/8	BBC	Studio monitoring loudspeakers with power amplifiers
1	NextStation	Next	Unix computer
1			Hand held control unit
2			RS232 to RS422 interface converters

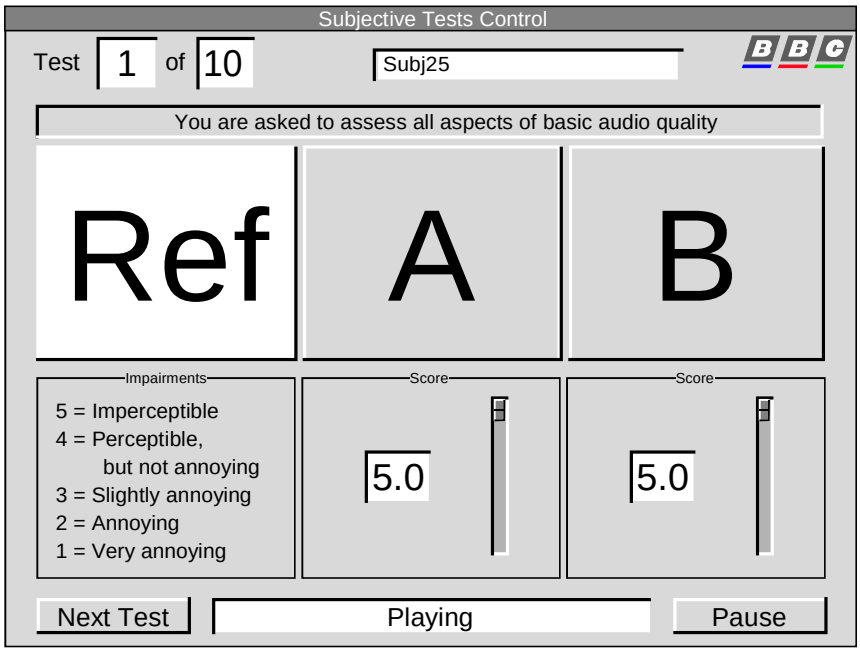
Arrangement of Equipment



Remote control keypad.
(used by the listener to control the test and enter grades).



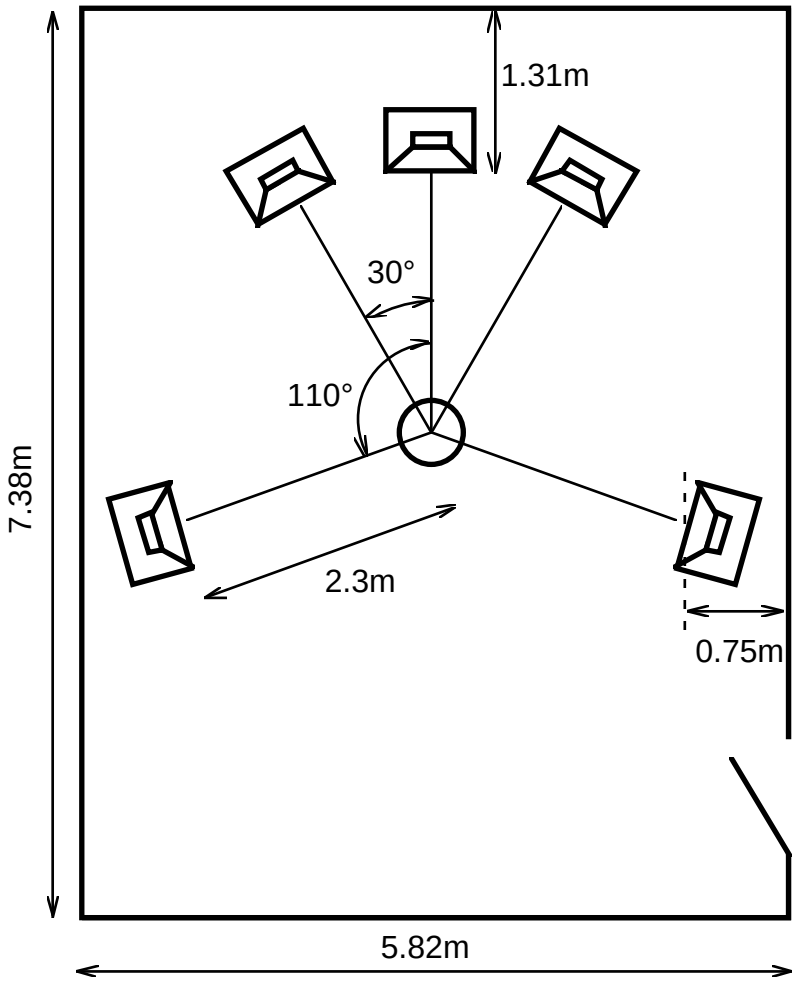
Control system display



Annex D: NHK test site: Listening Room Conditions and Equipment

Listening Room Conditions

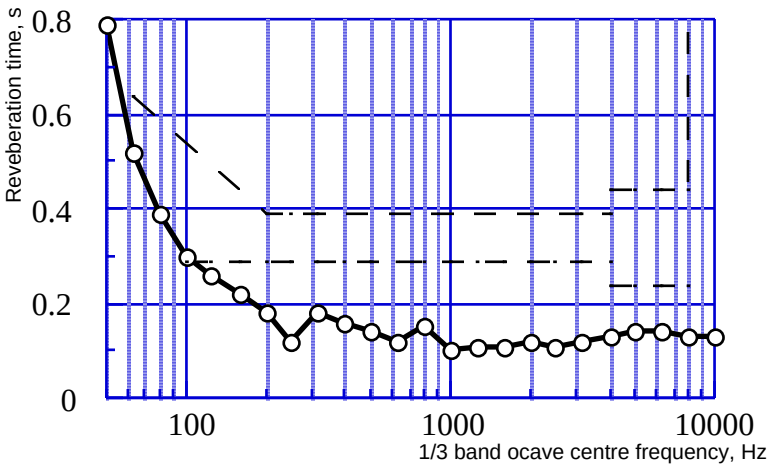
Room Dimensions



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

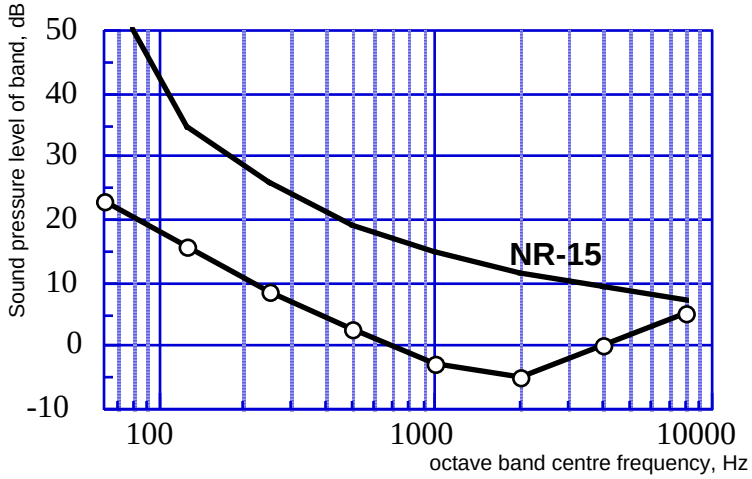
Listening Room B268, NHK Science & Technical Research Laboratories.

Reverberation time



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

Background noise

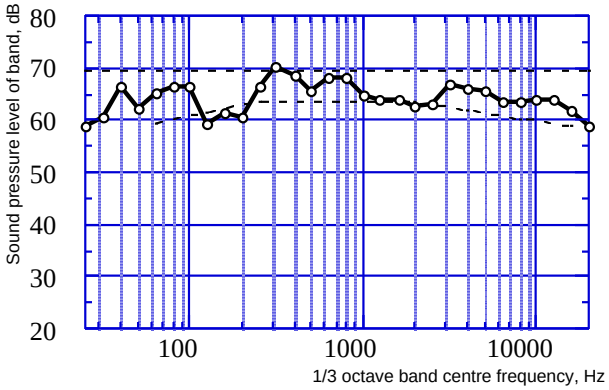


The noise level at the reference listening position meets the noise criterion NR-15.

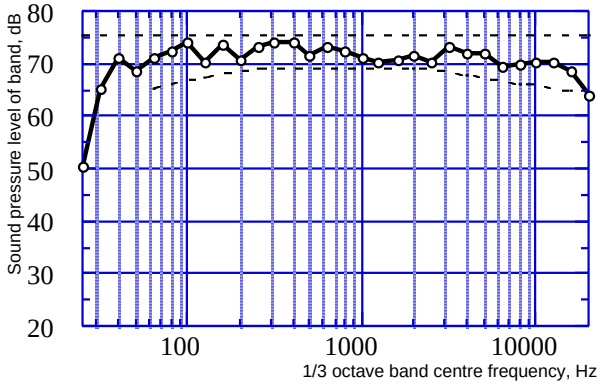
Frequency response measurements
(The listening tests were carried out without equalisation)

Left front

Without equalisation.

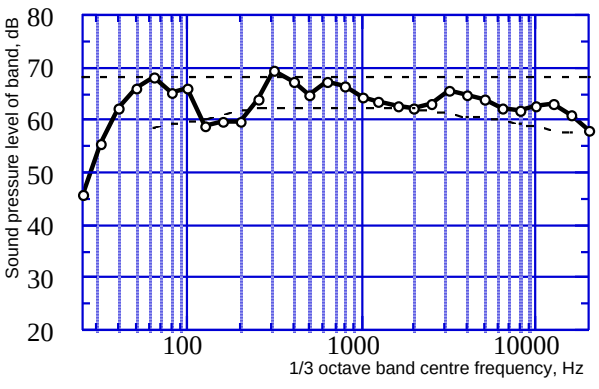


With equalisation.

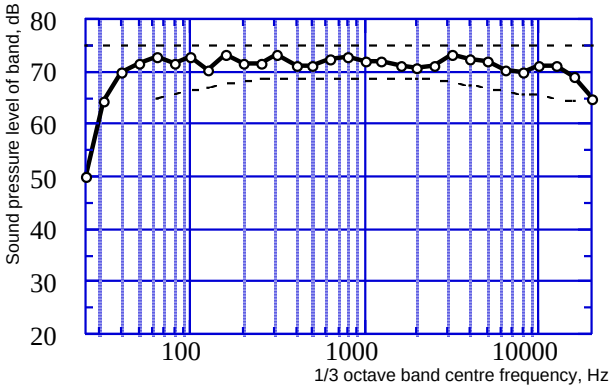


Centre front

Without equalisation.

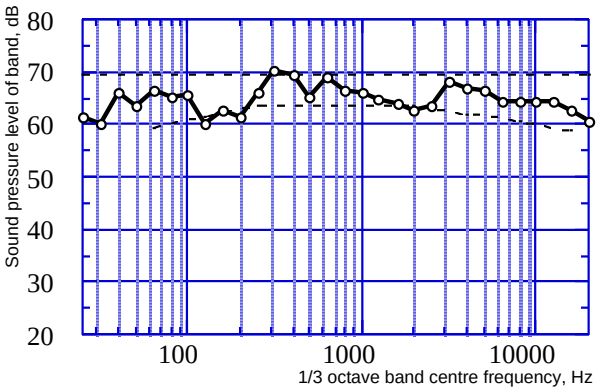


With equalisation

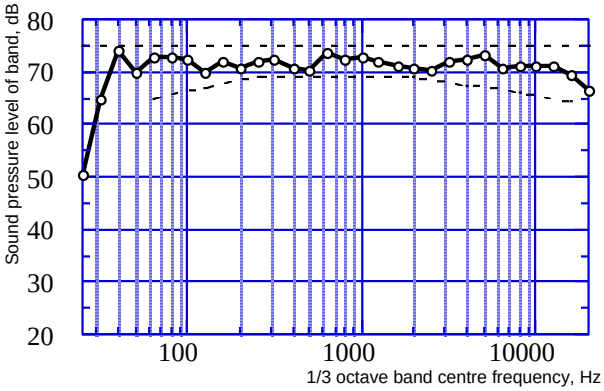


Right front

Without equalisation.

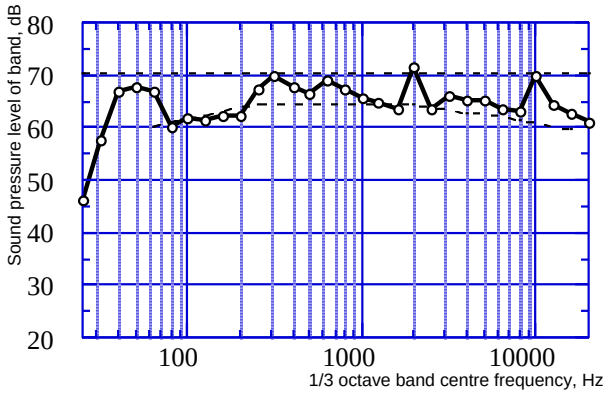


With equalisation

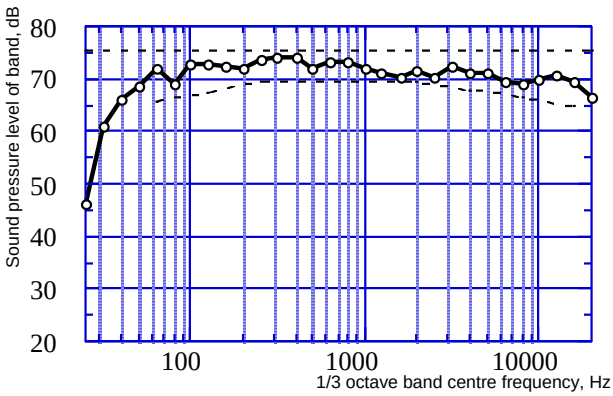


Left surround

Without equalisation

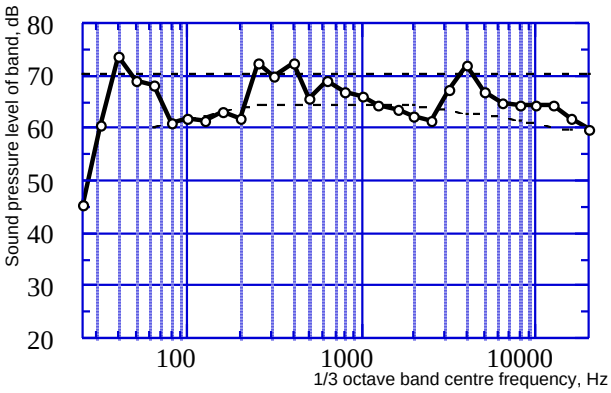


With equalisation

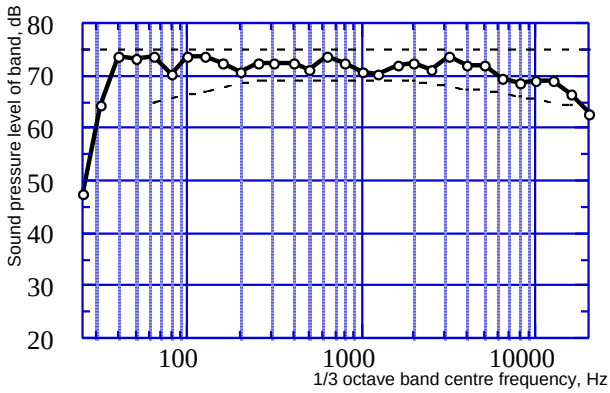


Right surround

Without equalisation



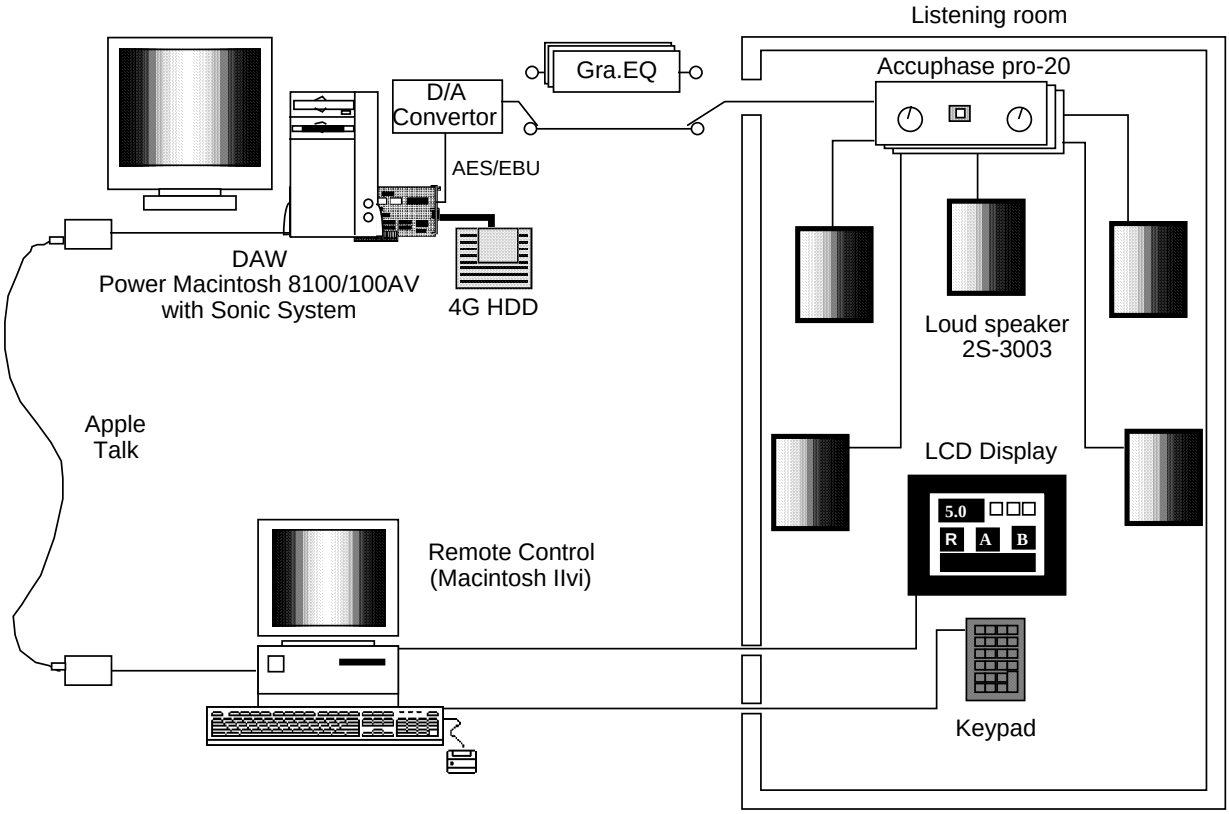
With equalisation



List of Test Equipment

Qty	Description	Model
1	Digital Audio Workstation	Sonic Studio Power Macintosh 8100/100AV
1	Remote Control Computer	Macintosh II vi
1	LCD	Sharp 9E-HC1
3	D/A Converter Unit	Yamaha DA2X
5	Loudspeaker Unit	Mitsubishi 2S-3003
3	Amplifier	Accuphase PRO-20
3	Equalizer	Yamaha DEQ5E

Playback System



Remote Control System

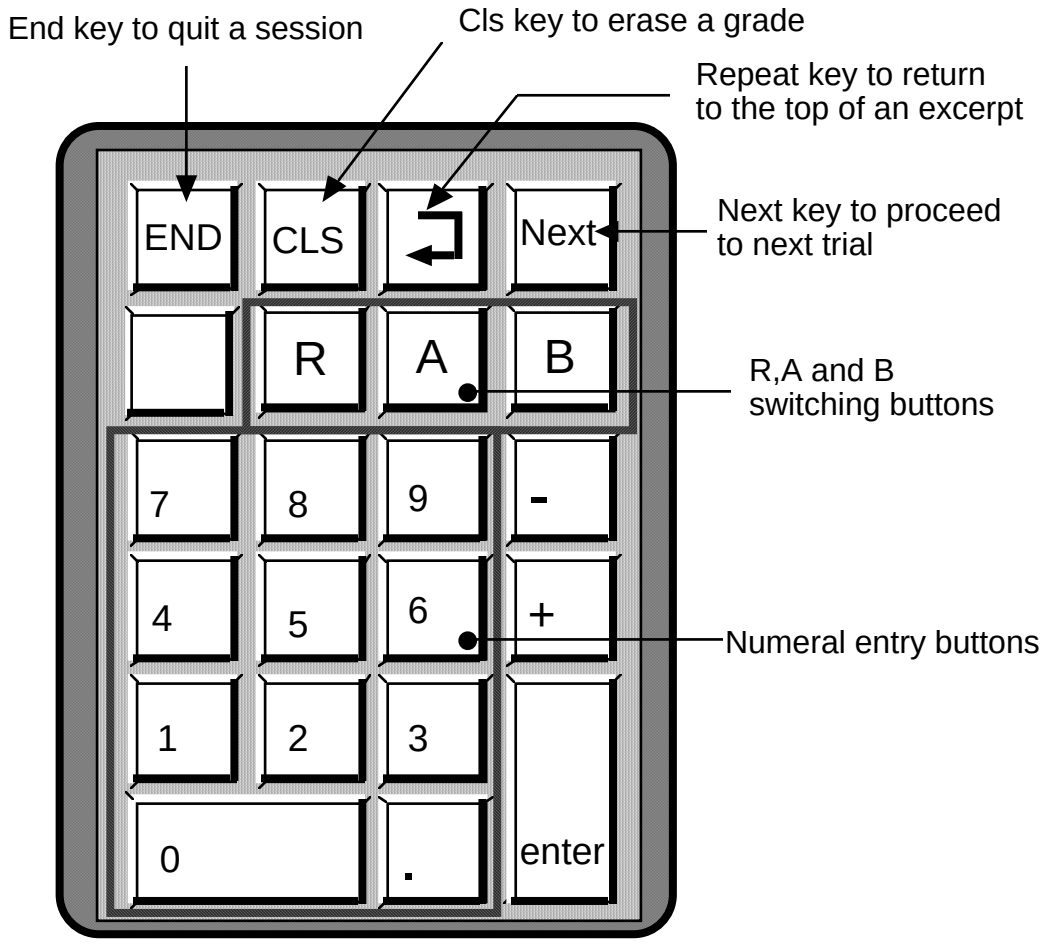


Figure 1 Sketch of keypad

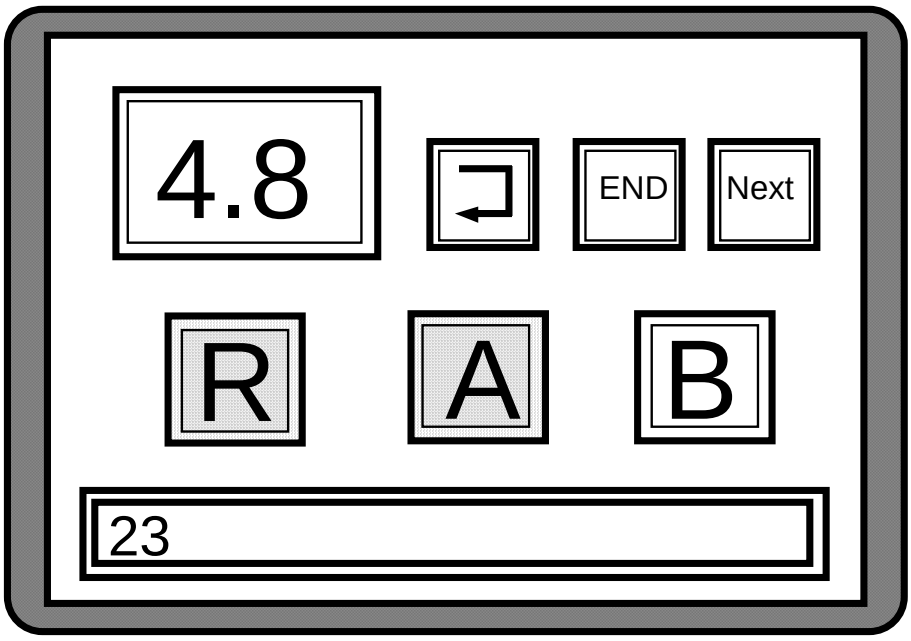


Figure 2 Sketch of Screen display

Annex E: Listeners participating in the tests

Listeners participating at the BBC.

Name	Organisation
Eric Beauchamp	BBC R&D Department
Ted de Bono	BBC Radio
Robin Cherry	BBC Radio
Adrian Chinnery	BBC Project Management Services
Paul Cunliffe	BBC Television
Nick Cutmore	BBC R&D Department
Tim Davies	BBC Television
Mike Felton	BBC Television
John Fletcher	BBC R&D Department
Rupert Flindt	BBC Radio
Curt Forsmark	Swedish Television
Nigel Gaylor	The Decca Recording Company
Neil Gilchrist	BBC R&D Department
Alice Grattan	BBC R&D Department
Jeff Hamilton	General Instrument, USA
Dave Hill	BBC Television
Toru Imai	NHK Sci. & Tech. Res. Labs., Japan
Howard Jones	BBC Television
Douglas McKinnie	University of Surrey, UK
Andrew McParland	BBC R&D Department
David Meares	BBC R&D Department
Lars Mossberg	Swedish Radio
Tony Philpott	BBC Television
Chris Poole	BBC R&D Department
Hugh Robjohns	BBC CBST
Florian Schmidt	Freelance Tonmeister, Berlin
Jonathan Smith	University of Surrey, UK
Ben deVille	University of Surrey, UK
Marvin Ware	BBC Radio
Richard West	BBC Television
David White	BBC R&D Department
Nick Zakarov	Nokia, Finland

Listeners participating at NHK.

Name	Organisation
Hiroyuki Fukuchi	Nippon Steel Corporation
Hiroshi Iriyama	Yamaha Corporation
Goro Tsutaya	Nippon Steel Corporation
Shigeki Fujii	Yamaha Corporation
Masami Suzuki	Pioneer Electronic Corporation
Kengo Nishimoto	NHK Broadcast Engineering Dept.
Itaru Kaneko	ASCII Corporation
Tatsuya Okada	Waseda University
Toru Shinmura	NHK Broadcast Engineering Dept.
Sadahiro Yasura	Victor Company of Japan, Limited
Takehiko Kuran	Victor Company of Japan, Limited
Mikihiko Okamoto	NHK Broadcast Engineering Dept.
Hiroyuki Okubo	NHK Sci. & Tech. Res. Labs.
Masamichi Otani	NHK Sci. & Tech. Res. Labs.
Takashi Katayama	Matsushita Electric Industrial Co., Ltd.
Masaichiro Maeda	Toshiba Corporation
Kanji Ohshima	NHK Broadcast Engineering Dept.
Yasushi Nakayama	NHK Sci. & Tech. Res. Labs.
Yasuji Ohta	Fujitsu Laboratories, Limited.
Kazuho Ono	NHK Sci. & Tech. Res. Labs.
Masakazu Iwaki	NHK Sci. & Tech. Res. Labs.
Teruji Kobayashi	Nittobo Acoustic Engineering Co., Ltd.
Jun Tamaru	NHK Broadcast Engineering Dept.
Kenichiro Masaoka	NHK Sci. & Tech. Res. Labs.

Annex F. Statistical Procedures

F.1 Tests for listener reliability

To assess the reliability of each listener, a one-sided t-test was employed to test the hypothesis that the mean value of all the diffgrades, i.e. (grade for the coded version) - (grade for the reference), was greater than zero. Rejection of the hypothesis, i.e. $p<0.05$, led to the conclusion that the listener was able to be accepted for analysis.

The Wilcoxon test was used as a confirmatory test. Had the inferences between the 2 tests differed for a large number of individuals, there would have been concern about the validity of the procedure. In fact, for only 1 out of the 56 subjects was there a disagreement, with this subject being rejected by the t-test but accepted by the Wilcoxon test. Further investigation of the diffgrades for this subject showed that there were high scores in both directions (i.e. -3.0 and +3.0) as well as many scores of zero. So the decision of the more sensitive t-test to reject the subject was upheld. It would have been of more concern had the less sensitive Wilcoxon test led to more subjects being rejected, since this would have indicated that a few large diffgrades might have exerted undue influence to save such subjects from being rejected by the t-test.

F.2 General comments on procedures

The 5% level of significance was used in all statistical tests.

Analysis of variance (ANOVA) was used for comparisons of means. To be consistent with previous reports [1, 4], the fixed effect model was used rather than the random effects model.

Levene's test was performed to determine homogeneity of variances.

The Shapiro-Wilk test was used to determine normality of data, it being more sensitive than the previously used Kolmogorov-Smirnov test.

F.3 Interpretation of the ANOVAs of NHK data

At first sight, the results from the full two-way ANOVA for NHK (Section 6.6.1), which show that the Codec main effect is not significant, may appear to contradict the results in Section 6.8.1 which reveal differences between MPEG-2 Layer II BC and MPEG-2 NBC at 320 kbit/s. An explanation of this point may therefore be worthwhile.

In the two-way ANOVA for NHK (Section 6.6.1), the main effect Codec has 3 degrees of freedom, one of which is the contrast between MPEG-2 Layer II BC and MPEG-2 NBC at 320 kbit/s, which was specified a priori. It is perfectly reasonable to examine separately any contrasts that have been chosen a priori, which is what has been performed in the Layer II / NBC(320) ANOVA. In the full two-way ANOVA, what has happened is that the main contrast of interest has contributed almost all of the codec main effects sum of squares (examination of the codec mean diffgrades

shows that Layer II and NBC(320) have the most extreme mean diffgrades). However, any other two orthogonal contrasts to this one will contribute very little to this sum of squares. So what has happened arithmetically is that Layer II / NBC(320) difference has been diluted by the apparent absence of other effects and this has led to a non-significant result (even so, the overall Codec main effect is "nearly" significant at 0.056). This effect can happen with F-tests with more than one degree of freedom. In the specific ANOVA for the MPEG-2 Layer II / NBC (320) comparison (Section 6.8.1), this dilution does not occur and main effect Codec is then found to have a significant influence.

Annex G. Numerical results

G.1 Summary of items effects: one-way ANOVAs

The following table summarises the one-way ANOVAs for each codec at each site and also gives the results of the Levene test in each case.

Source	Degrees of freedom	Sum of squares	Mean squares	Ratio	Prob.
Site: BBC. Codec: Layer II at 640 kbit/s					
Between Groups	9	58.5247	6.5027	5.3043	0.0000
Within Groups	220	269.7087	1.2259		
Total	229	328.2334			
Levene Test for Homogeneity of Variances					
Statistic	df1	df2	2-tail Sig.		
2.8744	9	220	0.003		
Site: BBC. Codec: NBC at 256 kbit/s					
Between Groups	9	17.2541	1.9171	1.4001	0.1893
Within Groups	220	301.2365	1.3693		
Total	229	318.4906			
Levene Test for Homogeneity of Variances					
Statistic	df1	df2	2-tail Sig.		
2.7280	9	220	0.005		
Site: BBC. Codec: NBC at 320 kbit/s					
Between Groups	9	8.5427	0.9492	1.1183	0.3508
Within Groups	220	186.7383	0.8488		
Total	229	195.2809			
Levene Test for Homogeneity of Variances					
Statistic	df1	df2	2-tail Sig.		
1.1019	9	220	0.362		
Site: BBC. Codec: NBC low complexity at 320 kbit/s					
Between Groups	9	22.5129	2.5014	2.5681	0.0079
Within Groups	220	214.2904	0.9740		
Total	229	236.8033			
Levene Test for Homogeneity of Variances					
Statistic	df1	df2	2-tail Sig.		
1.3841	9	220	0.197		
Site: NHK. Codec: Layer II at 640 kbit/s					
Between Groups	9	10.2488	1.1388	1.5449	0.1372
Within Groups	150	110.5681	0.7371		
Total	159	120.8169			
Levene Test for Homogeneity of Variances					

Statistic	df1	df2	2-tail Sig.		
1.5188	9	150	0.146		
Site: NHK. Codec: NBC at 256 kbit/s					
Between Groups	9	14.4413	1.6046	1.6157	0.1154
Within Groups	150	148.9706	0.9931		
Total	159	163.4119			
Levene Test for Homogeneity of Variances					
Statistic	df1	df2	2-tail Sig.		
0.9672	9	150	0.470		
Site: NHK. Codec: NBC at 320 kbit/s					
Between Groups	9	9.1240	1.0138	1.5630	0.1313
Within Groups	150	97.2937	0.6486		
Total	159	106.4177			
Levene Test for Homogeneity of Variances					
Statistic	df1	df2	2-tail Sig.		
1.1069	9	150	0.361		
Site: NHK. Codec: NBC low complexity at 320 kbit/s					
Between Groups	9	13.8188	1.5354	2.1734	0.0269
Within Groups	150	105.9706	0.7065		
Total	159	119.7894			
Levene Test for Homogeneity of Variances					
Statistic	df1	df2	2-tail Sig.		
0.8562	9	150	0.566		

G.2 Means and confidence intervals for each codec

G.2.1 BBC results.

Low anchor presentations: BBC results

(The confidence intervals have been calculated using the sample standard deviations).

Item	Mean	Standard deviation	95% Confidence interval: lower limit	95% Confidence interval: upper limit	Mean from MPEG '94 test
Harp	-2.213	1.337	-2.791	-1.635	-1.93
Manc	-1.361	1.681	-2.088	-0.634	-1.81
Pipe	-2.965	0.935	-3.369	-2.561	-2.25
Tria	-1.417	1.237	-1.952	-0.882	-1.22

Layer II at 640 kbit/s: BBC results

(The ANOVA assumptions fail and so the confidence intervals have been calculated using the sample standard deviations).

Item	Mean	Standard deviation	95% Confidence interval: lower limit	95% Confidence interval: upper limit
Cast	-0.426	1.251	-0.967	0.115
Clarinet	-0.235	0.800	-0.581	0.111
Eliot	-0.652	1.482	-1.293	-0.011
Glock	0.078	1.050	-0.376	0.532
Harp	-0.570	0.968	-0.989	-0.151
Manc	-0.357	0.922	-0.756	0.042
Pipe	-1.852	1.450	-2.479	-1.225
Station	-0.078	0.902	-0.468	0.312
Thal	-0.561	1.094	-1.034	-0.088
Tria	-0.222	0.923	-0.621	0.177

MPEG-2 NBC at 256 kbit/s: BBC results

(The ANOVA assumptions fail and so the confidence intervals have been calculated using the sample standard deviations).

Item	Mean	Standard deviation	95% Confidence interval: lower limit	95% Confidence interval: upper limit
Cast	-0.370	1.126	-0.857	0.117
Clarinet	-1.065	1.506	-1.716	-0.414
Eliot	-0.526	1.208	-1.048	-0.004
Glock	-0.604	1.185	-1.116	-0.092
Harp	-0.904	1.012	-1.342	-0.466
Manc	-0.683	1.071	-1.146	-0.220
Pipe	-1.030	1.155	-1.529	-0.531
Station	-1.052	1.504	-1.702	-0.402
Thal	-0.261	0.818	-0.615	0.093
Tria	-0.791	0.926	-1.191	-0.391

MPEG-2 NBC at 320 kbit/s: BBC results
(The ANOVA assumptions are valid and so the confidence intervals have been calculated using the ANOVA estimate of standard deviation).

Item	Mean	Error mean square	95% Confidence interval: lower limit	95% Confidence interval: upper limit
Cast	-0.404	0.8488	-0.781	-0.027
Clarinet	-0.678	0.8488	-1.055	-0.301
Eliot	-0.265	0.8488	-0.642	0.112
Glock	-0.139	0.8488	-0.516	0.238
Harp	-0.613	0.8488	-0.990	-0.236
Manc	-0.304	0.8488	-0.681	0.073
Pipe	-0.252	0.8488	-0.629	0.125
Station	-0.174	0.8488	-0.551	0.203
Thal	-0.065	0.8488	-0.442	0.312
Tria	-0.483	0.8488	-0.860	-0.106

MPEG-2 NBC low complexity at 320 kbit/s: BBC results
(The ANOVA assumptions are valid and so the confidence intervals have been calculated using the ANOVA estimate of standard deviation).

Item	Mean	Error mean square	95% Confidence interval: lower limit	95% Confidence interval: upper limit
Cast	-0.217	0.9740	-0.620	0.186
Clarinet	-0.265	0.9740	-0.668	0.138
Eliot	0.213	0.9740	-0.190	0.616
Glock	-0.335	0.9740	-0.738	0.068
Harp	-0.939	0.9740	-1.342	-0.536
Manc	-0.378	0.9740	-0.781	0.025
Pipe	-0.848	0.9740	-1.251	-0.445
Station	-0.170	0.9740	-0.573	0.233
Thal	-0.270	0.9740	-0.673	0.133
Tria	-0.409	0.9740	-0.812	-0.006

G.2.2 NHK results.

Low anchor presentations: NHK results
(The confidence intervals have been calculated using the sample standard deviations).

Item	Mean	Standard deviation	95% Confidence interval: lower limit	95% Confidence interval: upper limit
Harp	-1.581	1.387	-2.320	-0.842
Manc	-0.962	0.866	-1.423	-0.501
Pipe	-2.338	1.258	-3.008	-1.668
Tria	-0.769	0.805	-1.198	-0.340

Layer II at 640 kbit/s: NHK results

(The ANOVA assumptions are valid and so the confidence intervals have been calculated using the ANOVA estimate of standard deviation).

Item	Mean	Error mean square	95% Confidence interval: lower limit	95% Confidence interval: upper limit
Cast	-0.600	0.7371	-1.021	-0.179
Clarinet	-0.169	0.7371	-0.590	0.252
Eliot	-0.144	0.7371	-0.565	0.277
Glock	-0.419	0.7371	-0.840	0.002
Harp	-0.319	0.7371	-0.740	0.102
Manc	-0.144	0.7371	-0.565	0.277
Pipe	-1.038	0.7371	-1.459	-0.617
Station	-0.425	0.7371	-0.846	-0.004
Thal	-0.444	0.7371	-0.865	-0.023
Tria	-0.444	0.7371	-0.865	-0.023

MPEG-2 NBC at 256 kbit/s: NHK results

(The ANOVA assumptions are valid and so the confidence intervals have been calculated using the ANOVA estimate of standard deviation).

Item	Mean	Error mean square	95% Confidence interval: lower limit	95% Confidence interval: upper limit
Cast	-0.012	0.9931	-0.501	0.476
Clarinet	-0.187	0.9931	-0.677	0.301
Eliot	-0.263	0.9931	-0.752	0.226
Glock	-0.112	0.9931	-0.602	0.376
Harp	-0.413	0.9931	-0.902	0.076
Manc	-0.063	0.9931	-0.552	0.426
Pipe	-0.594	0.9931	-1.083	-0.105
Station	-0.994	0.9931	-1.483	-0.505
Thal	-0.063	0.9931	-0.552	0.426
Tria	-0.594	0.9931	-1.083	-0.105

MPEG-2 NBC at 320 kbit/s: NHK results

(The ANOVA assumptions are valid and so the confidence intervals have been calculated using the ANOVA estimate of standard deviation).

Item	Mean	Error mean square	95% Confidence interval: lower limit	95% Confidence interval: upper limit
Cast	-0.131	0.6486	-0.525	0.263
Clarinet	-0.069	0.6486	-0.463	0.325
Eliot	0.138	0.6486	-0.256	0.532
Glock	-0.494	0.6486	-0.888	-0.100
Harp	0.325	0.6486	-0.069	0.719
Manc	-0.300	0.6486	-0.694	0.094
Pipe	-0.481	0.6486	-0.875	-0.087
Station	-0.131	0.6486	-0.525	0.263
Thal	-0.200	0.6486	-0.594	0.194
Tria	-0.194	0.6486	-0.588	0.200

MPEG-2 NBC low complexity at 320 kbit/s: NHK results
(The ANOVA assumptions are valid and so the confidence intervals have been calculated using the ANOVA estimate of standard deviation).

Item	Mean	Error mean square	95% Confidence interval: lower limit	95% Confidence interval: upper limit
Cast	0.156	0.7065	-0.256	0.568
Clarinet	-0.275	0.7065	-0.687	0.137
Eliot	-0.138	0.7065	-0.550	0.274
Glock	-0.125	0.7065	-0.537	0.287
Harp	-0.663	0.7065	-1.075	-0.251
Manc	-0.088	0.7065	-0.500	0.324
Pipe	-0.825	0.7065	-1.237	-0.413
Station	0.075	0.7065	-0.337	0.487
Thal	-0.444	0.7065	-0.856	-0.032
Tria	-0.256	0.7065	-0.668	0.156

G.3 Comparison of MPEG-2 Layer II at 640 kbit/s and MPEG-2 NBC at 320 kbit/s

Note: Throughout these tabulations of differences, a positive mean implies that, for that item, the first named codec was awarded a better diffgrade than the second.

BBC results: Differences between diffgrades
MPEG-2 NBC at 320 kbit/s - Layer II at 640 kbit/s: 23 samples

Item	Mean	Standard deviation	Standard error	95% conf. int: lower limit	95% conf. int: upper limit
Cast	0.022	1.711	0.357	-0.718	0.762
Clarinet	-0.443	1.464	0.305	-1.077	0.19
Eliot	0.387	1.647	0.343	-0.325	1.099
Glock	-0.217	1.097	0.229	-0.692	0.257
Harp	-0.043	1.209	0.252	-0.567	0.48
Manc	0.052	0.863	0.180	-0.321	0.426
Pipe	1.6	1.626	0.339	0.897	2.303
Station	-0.096	1.113	0.232	-0.577	0.386
Thal	0.496	1.411	0.294	-0.115	1.106
Tria	-0.261	1.124	0.234	-0.747	0.225

NHK results: Differences of diffgrades.
MPEG-2 NBC at 320 kbits/s - Layer II at 640 kbits/s: 16 samples

Item	Mean	Standard deviation	Standard error	95% conf. int: lower limit	95% conf. int: upper limit
Cast	0.469	0.855	0.214	0.013	0.924
Clarinet	0.1	1.204	0.301	-0.542	0.742
Eliot	0.281	1.614	0.403	-0.579	1.141
Glock	-0.075	1.079	0.270	-0.65	0.5
Harp	0.644	1.249	0.312	-0.022	1.309
Manc	-0.156	0.968	0.242	-0.672	0.36
Pipe	0.556	0.751	0.188	0.156	0.957
Station	0.294	0.845	0.211	-0.157	0.744

Item	Mean	Standard deviation	Standard error	95% conf. int: lower limit	95% conf. int: upper limit
Thal	0.244	1.128	0.282	-0.358	0.845
Tria	0.25	0.897	0.224	-0.228	0.728

G.4 Comparison of MPEG-2 Layer II at 640 kbit/s and MPEG-2 NBC at 320 kbit/s

Note: Throughout these tabulations of differences, a positive mean implies that, for that item, the first named codec was awarded a better diffgrade than the second.

BBC results: Differences of diffgrades.
MPEG-2 NBC low complexity - MPEG-2 NBC at 320 kbits/s: 23 samples

Item	Mean	Standard deviation	Standard error	95% conf. int: lower limit	95% conf. int: upper limit
Cast	0.187	1.279	0.267	-0.366	0.74
Clarinet	0.413	1.268	0.264	-0.135	0.961
Eliot	0.478	1.097	0.229	0.004	0.953
Glock	-0.196	1.145	0.239	-0.691	0.3
Harp	-0.326	0.986	0.206	-0.753	0.1
Manc	-0.074	1.506	0.314	-0.725	0.578
Pipe	-0.596	1.216	0.254	-1.122	-0.07
Station	0.004	0.973	0.203	-0.417	0.425
Thal	-0.204	0.720	0.150	-0.516	0.107
Tria	0.074	0.982	0.205	-0.351	0.499

NHK results: Differences of diffgrades.
MPEG-2 NBC low complexity - MPEG-2 NBC at 320 kbits/s: 16 samples

Item	Mean	Standard deviation	Standard error	95% conf. int: lower limit	95% conf. int: upper limit
Cast	0.287	1.097	0.274	-0.297	0.872
Clarinet	-0.206	1.291	0.323	-0.894	0.482
Eliot	-0.275	1.408	0.352	-1.026	0.476
Glock	0.369	1.249	0.312	-0.297	1.035
Harp	-0.988	1.414	0.354	-1.741	-0.234
Manc	0.213	1.745	0.436	-0.718	1.143
Pipe	-0.344	0.560	0.140	-0.642	-0.045
Station	0.206	1.023	0.256	-0.339	0.751
Thal	-0.244	1.248	0.312	-0.909	0.421
Tria	-0.062	1.106	0.276	-0.652	0.527

G.5 Tests on model assumptions

The results of Levene's test for the Homogeneity of Variances have already been given in Section G.1.

To check the normality of the data, the Shapiro-Wilk test was performed for each item within each codec at each site, giving the following results (p-levels below 0.05 indicate a departure from normality):

Site : BBC	Normality p-value
Codec : Layer II at 640	0.0001
Codec : NBC at 256	0.0406
Codec : NBC at 320	0.0001
Codec : NBC lc at 320	0.0276

Site : NHK	Normality p-value
Codec : Layer II at 640	0.0095
Codec : NBC at 256	0.0001
Codec : NBC at 320	0.0143
Codec : NBC lc at 320	0.0085

Annex H. EBU definition of Indistinguishable quality

The following text is reproduced verbatim from a submission by the EBU to CCIR dated 24 October 1991. It is reproduced only because of the difficulty some readers may have in acquiring the original document.

Documents
CCIR Study Groups
Period 1990 - 1994

Delayed Contribution
Document TG10-2/3-E only
28 October 1991
Original: English

Received: 24 October 1991

Subject: Question 86/10
Keywords: Quality requirements, bit-rate reduction, "indistinguishable" quality

European Broadcasting Union (EBU)

BASIC AUDIO QUALITY REQUIREMENTS FOR DIGITAL AUDIO BIT-RATE REDUCTION SYSTEMS FOR BROADCAST EMISSION AND PRIMARY DISTRIBUTION

1. Introduction

In previous contributions, the EBU has proposed various requirements for digital audio bit-rate reduction systems. Experience has revealed certain difficulties in interpreting those requirements dealing specifically with the basic audio quality to be achieved.

The purpose of this contribution is to present these objectives in somewhat more detail in order to remove some ambiguities.

2. Basic audio quality

Systems may be tested subjectively in accordance with CCIR Recommendation 562 <CCIR, 1986-1990a>, using the five-grade impairment scale. The double-stimulus method mentioned in this Recommendation is often favoured by broadcasters for subjective testing. However, a particularly advantageous method may be the double-blind test with a hidden reference <2>[6]. In this method, the subject has three signals, A, B and C. A is always the reference (unprocessed) source signal. The selection of the reference or system under test for presentation as B or C varies, and is not known to the subjects. The subjects have to grade both B and C, using a continuous impairment scale incorporating the five grades of the CCIR impairment scale.

The quality of the audio signals reproduced after decoding should be indistinguishable from the quality obtained from a compact disc. In practice, this implies comparing the quality of the analogue output of the codec (the input interface having a sampling frequency of 48 kHz) with the signal replayed from 16-bit linear equipment (also having a sampling frequency of 48 kHz). This should ideally apply for all types of programme material. A more complete explanation of practical target requirements, together with a mathematical basis for

[6] This is also known as the "Triple-stimulus hidden reference impairment method".

assessing the results of tests, is given in Annex 1. Using the double-blind test with a hidden reference, the mean grades must be consistently higher than 4 on the CCIR impairment scale. In any case, no test item may have a mean grade lower than 4.0.

It is inevitable that the source signal will be awarded a grade lower than 5 in such subjective tests. This does not justify any normalisation of the results for the systems under test.

Annex 1

Target requirements for the audio quality

As indicated in the main text, the quality of the signals after decoding must be "indistinguishable" from compact disc quality. In the present context, the term "indistinguishable" is defined as follows.

A decoded test item is indistinguishable from the reference test item when, using the triple stimulus hidden reference method, the 95 % confidence intervals of the test and reference subjective assessments overlap. The calculation of the confidence intervals relates to the population of each test item.

When applying the above rule, the subjective assessments must be made with not less than 40 subjects and not less than 8 test items, the latter being critical examples of normal broadcast material.

Ideally, all test items should be indistinguishable from the reference. In practice, however, reasonable criteria are at present as follows:

1. The results of at least 70 % of the total number of test items must overlap. That is, the upper limit of the coded signal confidence interval is greater than the lower limit of the reference signal confidence interval.

2. The remaining test items (up to 30 % of the total number) need not overlap, but must meet the following requirement.

   The ratio (upper limit of the test confidence interval) / (upper limit of the reference confidence interval) should be greater than or equal to 0.85.