



DRM Source Coding Group

Report on Subjective Listening Tests of SBR-LC, an AAC-based Audio Bandwidth Widening Tool

Source: BBC Research & Development, Kingswood Warren, UK.
Bosch, Hildesheim, Germany.
Deutsche Telekom T-Nova, Berlin, Germany.

Date: February 2001

List of authors (in alphabetical order):

Thomas Buchholz (T-Nova)
Torsten Mlasko, Frank Hofmann (Bosch)
Andrew Murphy (BBC)



1. Introduction

The DRM (Digital Radio Mondiale) project has made a choice as far as the audio coding technique is concerned: the MPEG-4 AAC low complexity profile is chosen and represents today the state of the art coding technique for audio signals at bit rates from 12 kbit/s to 64 kbit/s for a monophonic channel; MPEG-4 AAC has already been tested within the MPEG-4 standardisation body [1].

AAC proves to reach a very high subjective audio quality (i.e. close to transparent quality) at 56 kbit/s and above for a mono signal. However the DRM technique, where radio frequency resource is limited, is not able to convey such high bit rates. The available bit rate will be less than or equal to 24 kbit/s, which would not allow AAC to provide a “near CD-like” quality.

Nevertheless, audio bandwidth widening techniques can be used to enhance the perceived quality of the AAC coded signal at these low bit rates.

These listening tests examine one such system, SBR-LC, from Coding Technologies Sweden (CTS).

The questions to be addressed are as follows:

1. To evaluate the quality improvement reached by AAC + bandwidth widening methods at low bitrate in a new DRM system; this evaluation calls for the comparison of MPEG-4-AAC with and without bandwidth widening methods at the same bit rate
2. To evaluate the amount of saved bit rate that can be reached when using bandwidth widening methods; this evaluation calls for adding MPEG-4-AAC anchors at higher bit rates

The tests were carried out at three locations with independent outcomes. The locations were:

BBC R & D Department
Kingswood Warren, Tadworth, UK.

T-Nova Berkom
Goslarer Ufer 35, D-10589 Berlin, Germany.

Robert Bosch GmbH
Advanced Development Multimedia Systems, D-31132 Hildesheim, Germany.

At both Bosch and T-Nova, the listening tests made use of headphones, whilst at the BBC a single, centrally located loudspeaker was used.



2. Audio Codecs Under Test

There were four codecs under test as shown in *Table 1*. All were designed to code a single, mono channel of audio. All AAC implementations use 1024 FFT and no error resilience.

	Codec	Proponent	Input signal sampling rate	Total bitrate	Remarks
1	AAC Pure	CTS	48kHz	24 kbit/s	MPEG-4 AAC Codec
2	AAC SBR	CTS	48kHz	24 kbit/s	As Codec 1 including SBR-LC (the SBR extension operating at a data rate of between 2.5 and 3 kbps)
3	AAC SBR Core	CTS	48kHz	24 kbit/s	AAC Core of Codec 2
4	AAC Wideband	CTS	48kHz	32 kbit/s	MPEG-4 AAC Codec

Table 1 – The Audio Codecs Under Test

3. Test Material

3.1 Selection of Items

A total number of 10 critical test items were used in the test. These items had been selected from available test material used in former listening tests or DRM demonstrations [1, 2]. T-Nova had prepared the material and sent out a CD-ROM with 46 items to the codec provider. The length of each test item did not exceed 20 seconds. Coding was performed by the codec provider using the provided materials. All items were mono at a sample rate of 48 kHz. The decoded files as well as the bitstreams were uploaded to the T-Nova ftp site. All decoded audio files had a sampling rate of 48 kHz.

The final selection of ten appropriate test items was delegated to a selection panel, which consisted of three expert listeners from BBC R&D, Bosch and T-Nova. The selection process took place in Berlin at T-Nova.

The instructions to the selection panel were:

- select in total 10 critical items of typical broadcast material
- the material should include speech only, music only and speech together with (background) music
- find critical material for each of the coders (clearly audible impairments & different characteristics)
- to check if there is a difference (clearly audible impairments) between loudspeaker or headphone reproduction and report it
- select 3 or 4 suitable items used for training of the test persons



The following 10 items of typical broadcast material were found to be critical for all of the codecs under test by the selection panel.

No.	Item number	Name	Category
1	Item_01	Male voice (English)	Speech only
2	Item_06	Folk music	Music only
3	Item_12	Speech + noise (Swedish)	Speech + Music
4	Item_21	Susan Vega, Tom's diner	Speech only
5	Item_27	Complex (sound+applause)	Music only
6	Item_29	"route 66"	Music only
7	Item_37	mainly speech (Spanish news) 1	Speech + Music
8	Item_38	mainly speech (Spanish news) 2	Speech + Music
9	Item_40	speech only (English feature)	Speech only
10	Item_45	music, speech different languages	Music only

The following four items are recommended for training of the test subjects.

No.	Item number	Name	Category
1	Item_04	Speech + music	Speech + Music
2	Item_22	Speech (Male German speech)	Speech only
3	Item_23	Complex (Tracy Chapmann)	Music only
4	Item_26	Music (Palmtop boogie)	Music only

For more information see Annex A.

3.2 Verification

The codec proponent provided a software encoder and decoder, and this was used to verify the supplied .wav files. In addition, the bitstreams were checked and found to be within the specified bit-rates.

4. Experimental Design

4.1 Test Method

The test procedure followed that of the "Multiple Stimulus with Hidden Reference and Anchors" (MUSHRA) [3] method for the subjective assessment of intermediate quality audio.

The subject was presented with a series of trials, each corresponding to a different one of the audio items selected for the tests. In each trial, the subject was presented with the known reference version as well as a set of signals to be graded. The set of signals consisted of the four coding systems under test, two hidden anchors and a hidden copy of the reference, making a total of seven signals to be graded for each trial. The hidden anchors were bandwidth limited versions of the unprocessed, reference signal and were chosen as 3.5kHz and 7kHz for these tests.

Since the subjects can directly compare the impaired signals, this method provides the benefits of a full paired comparison test in that the subject can more easily detect differences between the impaired signals and grade them accordingly. This feature permits a high degree of resolution in the grades given to the systems. It is important to note, however, that subjects

derived their grade for a given system by comparing that system to the reference signal, as well as to the other signals in each trial.

The subjects were asked to grade *basic audio quality*, a single global attribute used to judge any and all detected differences between the reference and the object in question. The scoring is done according to the Continuous Quality Scale (CQS). The CQS consists of a continuous scale, labelled with five adjectives from top to bottom:

- Excellent
- Good
- Fair
- Poor
- Bad

The BBC and T-Nova test sites made use of the Canadian Research Centre's, System for the Evaluation of Audio Quality (CRC-SEAQ) and a typical screen shot of the computer-controlled replay system is shown in *Figure 3*. At Bosch, a system provided by the FhG was used. A typical screenshot is depicted in *Figure 4* (overleaf). The set of signals to grade were shown on buttons A to G along with the known reference. The grading scale was continuous from 0 to 100 in unit steps and grades are recorded by adjusting the slider corresponding to each to button. It was possible to switch cleanly between all of the signals at will even whilst they were playing. Additionally the subject was permitted to use looped playback and to focus in on a particular section of audio if required.

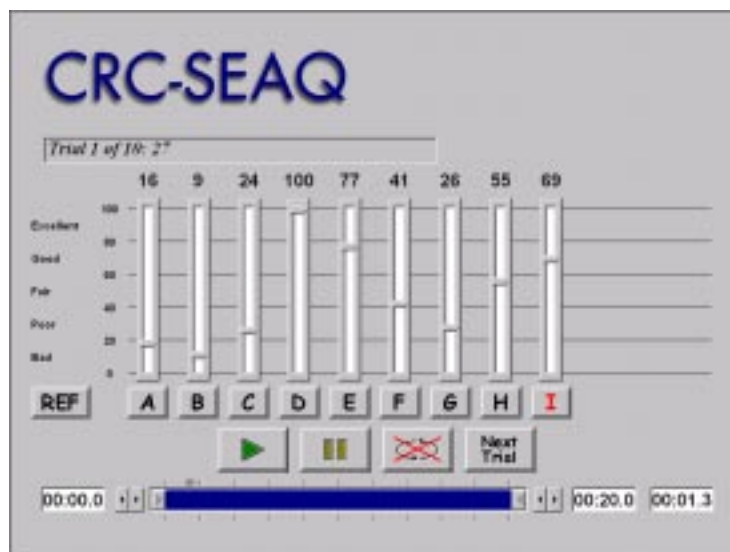


Figure 3 – The Computer-Controlled Replay System Used for the Tests (CRC-SEAQ) at the BBC and T-Nova Test Sites

The order of presentation of the trials and the allocation of the signals to be graded to the buttons A to G was randomised for every subject that took part.

The MUSHRA specification states that the acoustic environment should meet Sections 7 and 8 of ITU-R BS.1116-1 [4].



Figure 4 – The Computer-Controlled Replay System Used for the Tests at the Bosch Test Site (Fraunhofer Institute MUSHRA)

4.3 Training phase

In order to ensure that reliable results were obtained, it was necessary to train the subjects in special training sessions. The four training items that had been specially selected for these tests were used and the session was led by the report author at each site.

In preparation for the true evaluation phase, the training phase allowed the subjects to achieve two objectives as follows:

- (a) PART A – to become familiar with all the sound excerpts under test and their quality level ranges; and
- (b) PART B – to learn how to use the equipment and the grading scale.

In PART A, small groups of subjects were able to listen to examples of coded versions of the training items to gain a feel for the range of qualities from the different coding systems. The subjects were encouraged to discuss any audible artefacts.

In PART B, subjects were asked to use the scoring equipment individually to evaluate the quality of the four training items in one continuous session in terms of the CQS grading system.

The subjects were instructed that they should not necessarily give the grade “Bad” to the sound excerpt with the lowest quality in the test. However, one or more excerpts must be given the grade “Excellent” because the unprocessed reference is included as one of the excerpts to be graded.

No grades given during the training phase were taken into account in the true tests and the subjects were made aware of this.



4.4 Grading phase

For the grading phase, the subjects listened to the 10 audio items chosen by the selection panel in two equal sessions, separated by a break in order to avoid listener fatigue. The subjects were obliged to evaluate each session without stopping. The order of presentation of the items was randomised across both sessions.

4.5 Post-screening of subjects

As part of the general analysis two post-screening methods according to the MUSHRA procedure were used:

- one was based on the ability of the subject to make consistent repeated gradings; i.e for the hidden reference and the hidden anchors
- the other relied on inconsistencies of an individual grading compared with the mean result of all subjects for a given item.

The screening was carried out by looking to the individual spread and the deviation from the mean grading of all subjects at a given test site. The aim of this was to get a fair assessment of the quality of the test items. Due to the fact that “Intermediate Quality” was tested, a subject should have been able to identify the coded version very easily and therefore should have been able to give a grade that is in the range of the majority of the subjects. Subjects with grades at the upper end of the scale were likely to be less critical and subjects who have grades only at the lowest end of the scale were likely to be too critical.

The methods were primarily used to eliminate subjects who could not make the appropriate discriminations.



5. Test Site Details

5.1 BBC Test Site

Low-pass Filtering for the Anchors

For the production of the low-pass filtered anchors, *CoolEdit Pro* was used with the following settings:

- Butterworth Filter
- Cut-off Frequency 3500 Hz or 7000 Hz (as appropriate)
- 40th order

Time Alignment

Custom BBC software was used to detect the offset in samples between the coded wav files and the reference versions. *CoolEdit Pro* was then used to perform the time alignment on the coded audio files, inserting silence or removing samples from the beginning of the files as required.

Listening Panel

The listening panel at the BBC consisted of 22 subjects with a mixture of experts and non-experts taking part. Most of the listeners had a technical background and some were particularly involved in audio research.

Grading Phase

The grading phase consisted of one session containing all 10 items. Listeners were obliged, however, to perform the session in two continuous halves, taking a break after the first five items before returning to complete the session. This method ensured that the presentation of the items was completely randomised across the duration of the listening test.

Test Duration

The total duration for the tests varied between 2 hours for the quickest to 3½ hours of listening time for the slowest subject. This included training time of approximately 1 hour in total for parts A and B together.

Listening Conditions Including Listening Test System

Listening Room 3 at BBC Research & Development, Kingswood Warren, was used for the tests. The detailed characteristics of the room are given in Annex E and it meets the requirements of ITU-R BS.1116-1.

The playback system utilised a digital I/O card in the PC with an external DAC, the output of which was fed to a single, centrally located, high quality monitoring loudspeaker.



5.2 Bosch Test Site

Low-Pass Filtering for the Anchors

The anchors were provided by T-Nova. For the production of the low pass filtered anchors at T-Nova the following adjustments in CoolEdit Pro were used:

- Butterworth Filter
- Cut-off Frequency 3500 Hz, (7000 Hz)
- 40th order

Time Alignment

The time alignment was performed by T-Nova:

The decoded files were time aligned to the reference file by using the CRC TimeSync Software V1.1. After this procedure, all files (references, anchors and decoded files) were exactly synchronous and had the same length.

Grading Phase

For each of the test items, the signals under test were randomly assigned to the buttons. In addition, the test items were randomised for each subject within a session. To avoid sequential effects, each subject was running the two sessions in random order.

Listening Panel

The listening panel at Bosch consisted of 15 subjects. All listeners are male and in the age of 23 to 37. Some members of the listening panel, participated in MPEG tests before. All of the subjects have technical background. A few of them are also involved in audio matters.

Test Duration

The test phase for any listener consisted of a total time typically 2 to 2½ hours. The training phase was included in this time scheduled, with about ½ hour for training and instructions how to handle the equipment and how to use the scale. The subjects were obliged to evaluate each session without any break. The test for one subject was done within one day.

Listening Conditions Including Listening Test System

The tests were conducted using Sennheiser HD 545 headphones. The subjects had the possibility to set the reproduction level individually before they started the actual tests. The test was performed by one listener only at the time. The subjects were not allowed to change the reproduction level during the test.

The test items were stored on a Windows NT workstation which had a digital sound board connected to an external DAC. A specially designed software IAQ (FhG-MUSHRA) was running on the PC.



5.3 T-Nova Test Site

Low-Pass Filtering for the Anchors

For the production of the low-pass filtered anchors the following adjustments in CoolEdit Pro were used:

- Butterworth Filter
- Cut-off Frequency 3500 Hz, (7000 Hz)
- 40th order

Time Alignment

The decoded files were time aligned to the reference file by using the CRC TimeSync Software V1.1. After this procedure, all files (references, anchors and decoded files) were exactly synchronous and had the same length.

Grading Phase

The test was divided into two sessions, each with five trials. For each of the test items, the signals under test were randomly assigned to the buttons. In addition, the test items were randomised for each subject within a session. To avoid sequential effects, each subject was running the two sessions in random order.

Listening Panel

The listening panel at T-Nova consisted of 24 subjects, 5 women and 19 men, aged between 21 and 61 years. More than half of the panel had experience from previous listening tests.

Test Duration

The test phase for any listener consisted of a total time typically 2 to 3 hours with a minimum of ½ hour per session up to 1 hour per session. The training phase was included in this time scheduled, with about ½ hour for training and instructions how to handle the equipment and how to use the scale. The subjects were obliged to evaluate each session without any break. The test for one subject was done within one day.



Listening Conditions Including Listening Test System

The tests were conducted in the listening room 1 (compliant to ITU-BS.1116) at T-Nova in Berlin. Headphones were used as reproduction devices (Beyerdynamic DT-990). No loudspeaker listening was used. The subjects had the possibility to set the reproduction level individually before they started the actual tests. Up to three listeners were sitting in the same room separated by acoustic absorption walls.

The subjects were not restricted from changing the reproduction level during the test, however it is very unlikely that a subject changed the level during a session.

The test items were stored on a Windows 98 workstation which had a digital sound board (Creamware Pulsar) to play the audio to the input of the D/A-converter (Tascam DA-30) and after that to the headphones. Specially designed software IAQ (CRC-SEAQ) was running on the PC's.



5. Statistical Analysis

The statistical analysis followed standard MUSHRA procedure¹.

The raw data from the CRC-SEAQ software was a normalised score from 0 (bad quality) to 100 (excellent quality) for each item and codec for individual listeners. The calculation of the averages of normalised scores of all listeners remaining after post-screening will result in the Mean Subjective Scores (MSS).

The first step of the analysis of the results is the calculation of the mean score, \bar{u}_{jk} for each of the presentations:

$$\bar{u}_{jk} = \frac{1}{N} \sum_{i=1}^N u_{ijk} \quad (1)$$

where:

u_i = score of observer i for a given test condition j and sequence k
 N = number of observers

Confidence intervals were also calculated which were derived from the standard deviation and the size of each sample. The 95% confidence interval is given by:

$$[\bar{u}_{jk} - \delta_{jk}, \bar{u}_{jk} + \delta_{jk}]$$

where:

$$\delta_{jk} = 1.96 \frac{S_{jkl}}{\sqrt{N}} \quad (2)$$

and the standard deviation S_{jk} is given by:

$$S_{jk} = \sqrt{\sum_{i=1}^N \frac{(\bar{u}_{jk} - u_{ijk})^2}{(N-1)}} \quad (3)$$

With a probability of 95%, the absolute value of the difference between the experimental mean score and the “true” mean score (for a very high number of observers) is smaller than the 95% confidence interval, on condition that the distribution of the individual scores meets certain requirements.

Similarly, a standard deviation S_j could be calculated for each test condition. It is noted however that this standard deviation will, in cases where a small number of test sequences are used, be influenced more by differences between the test sequences used than by variations between the assessors participating in the assessment.

¹ At the time of writing the MUSHRA specification was a draft and recommended that confidence intervals should be calculated as described. The latest version, however, recommends the use of a Student-t distribution.



6. Results

6.1 General

The graphs show the statistical analysis of the previous section performed on the raw listener results.

Graphs 1, 5 & 9 show the mean and 95% confidence intervals for each item and codec individually. Separate graphs for each codec are reproduced in Annex B. *Graphs 2, 6 & 10* show the mean results for each item with each codec being a different series. *Graphs 3, 7 & 11* give mean and 95% confidence intervals for AAC SBR.

Finally, *Graphs 4, 8 & 12* give mean and 95% confidence intervals for each coding system overall. This has been produced by performing the statistical analysis on every listener and item for each codec.

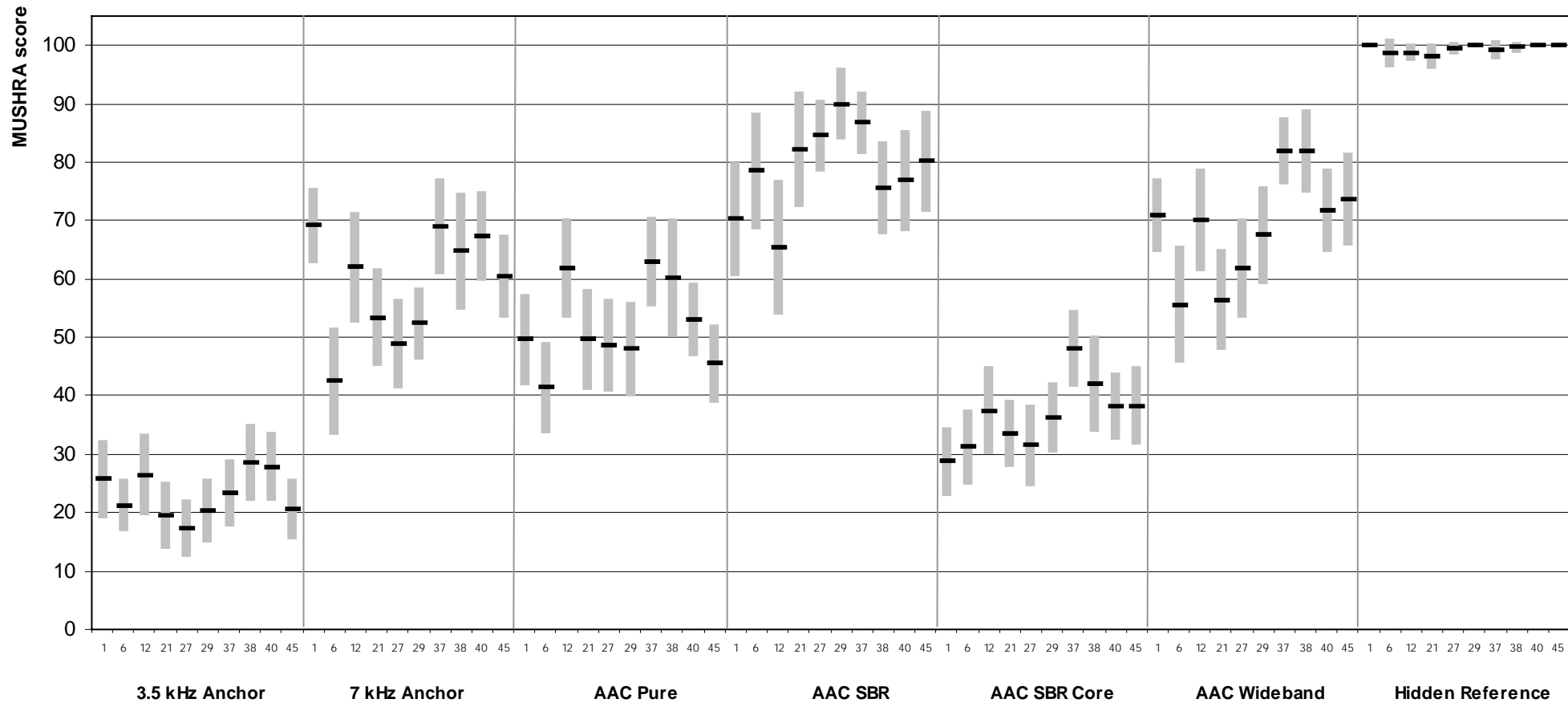
6.2 Post-screening

Post screening was carried out at the T-Nova test site, resulting in only two subjects having to be removed before the final evaluation. Both subjects evaluated, for example, the 7 kHz bandwidth-limited anchor and/or one or two codec versions clearly higher than the hidden reference for several items.



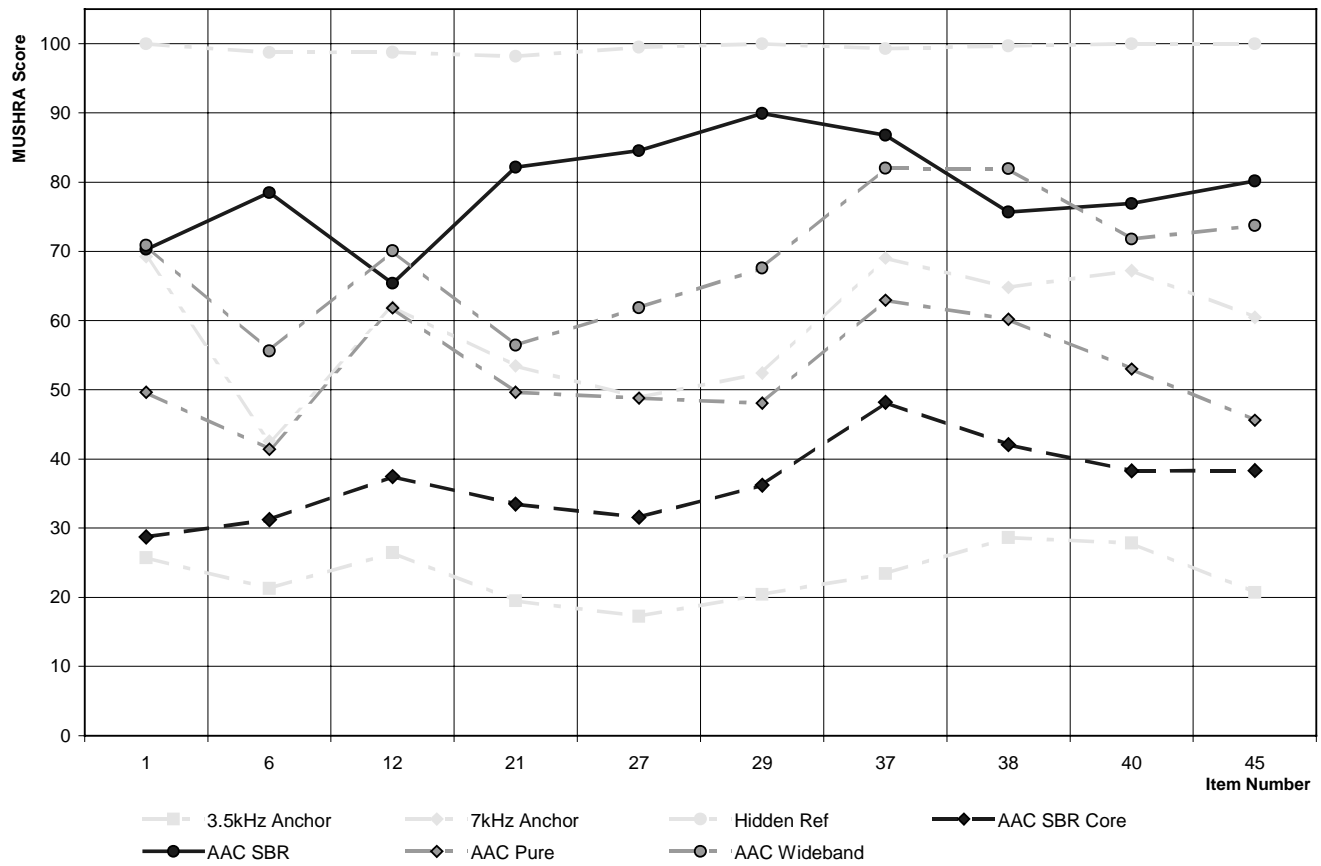
6.3 BBC Test Site

Graph 1 – BBC Test Site: Mean and 95% Confidence Intervals for All Items and Codecs

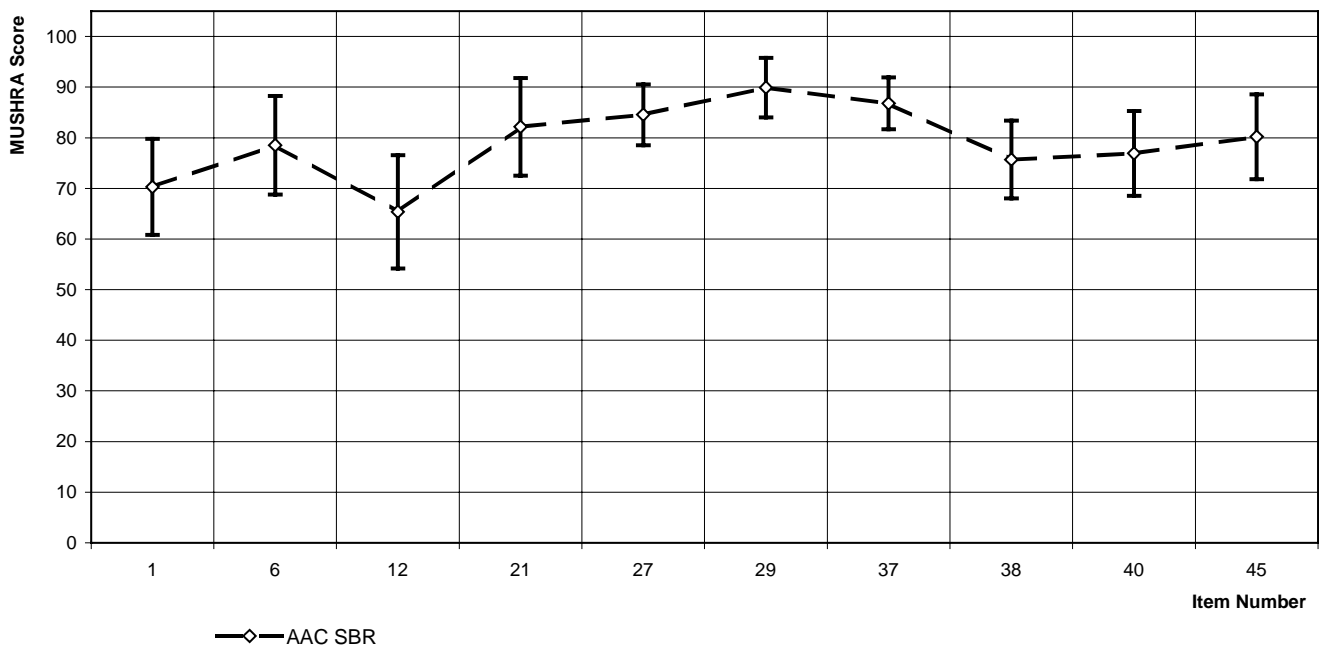




Graph 2 – BBC Test Site: Means for All Items and Codecs

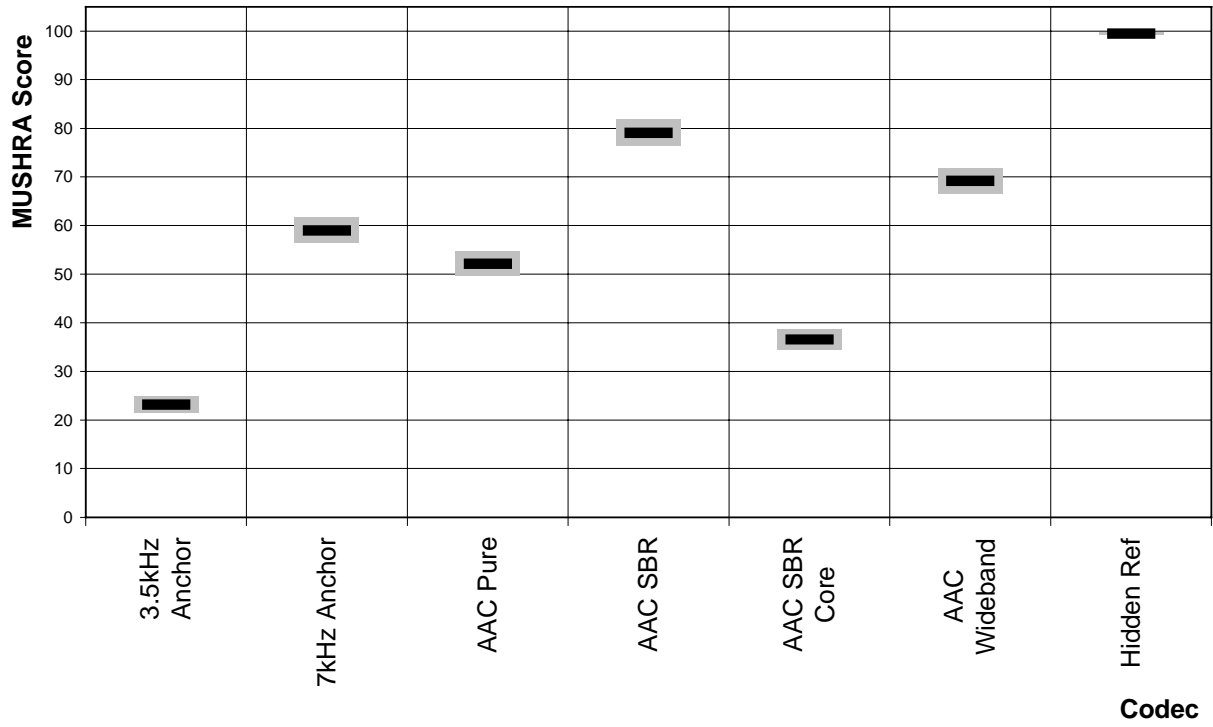


Graph 3 – BBC Test Site: Mean and 95% Confidence Intervals for AAC SBR





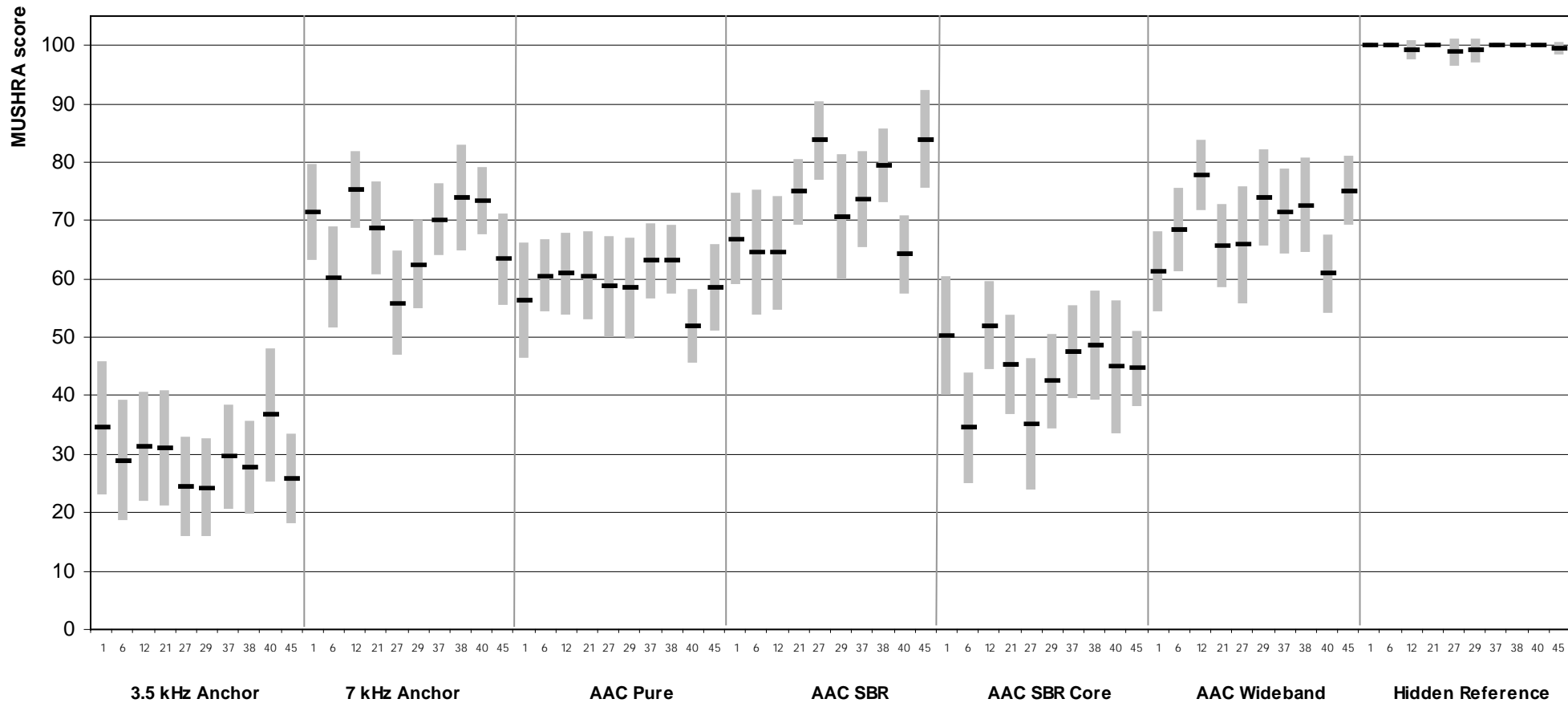
Graph 4 – BBC Test Site: Mean and 95% Confidence Intervals for each Codec Over All Items





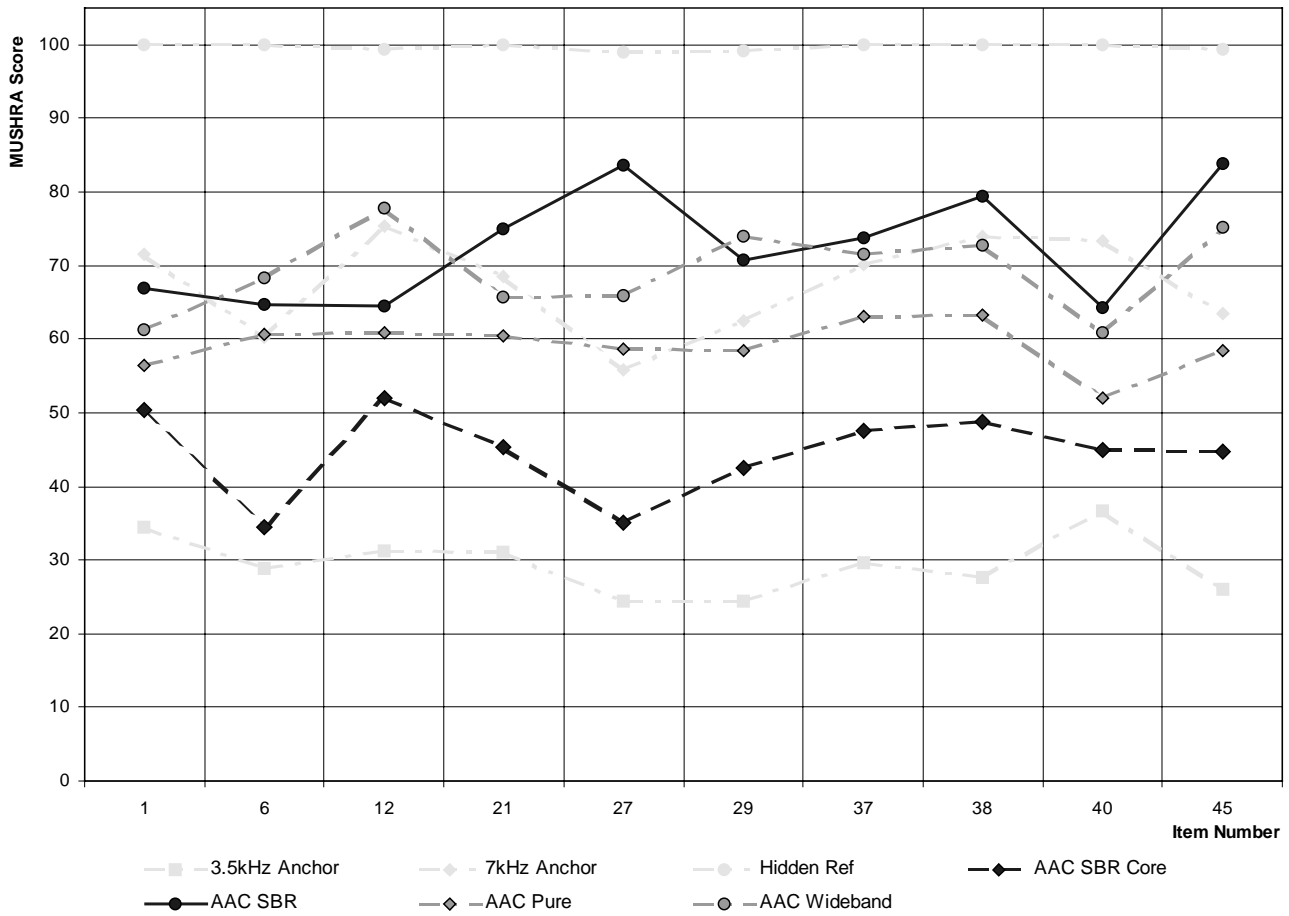
6.4 Bosch Test Site

Graph 5 – Bosch Test Site: Mean and 95% Confidence Intervals for All Items and Codecs

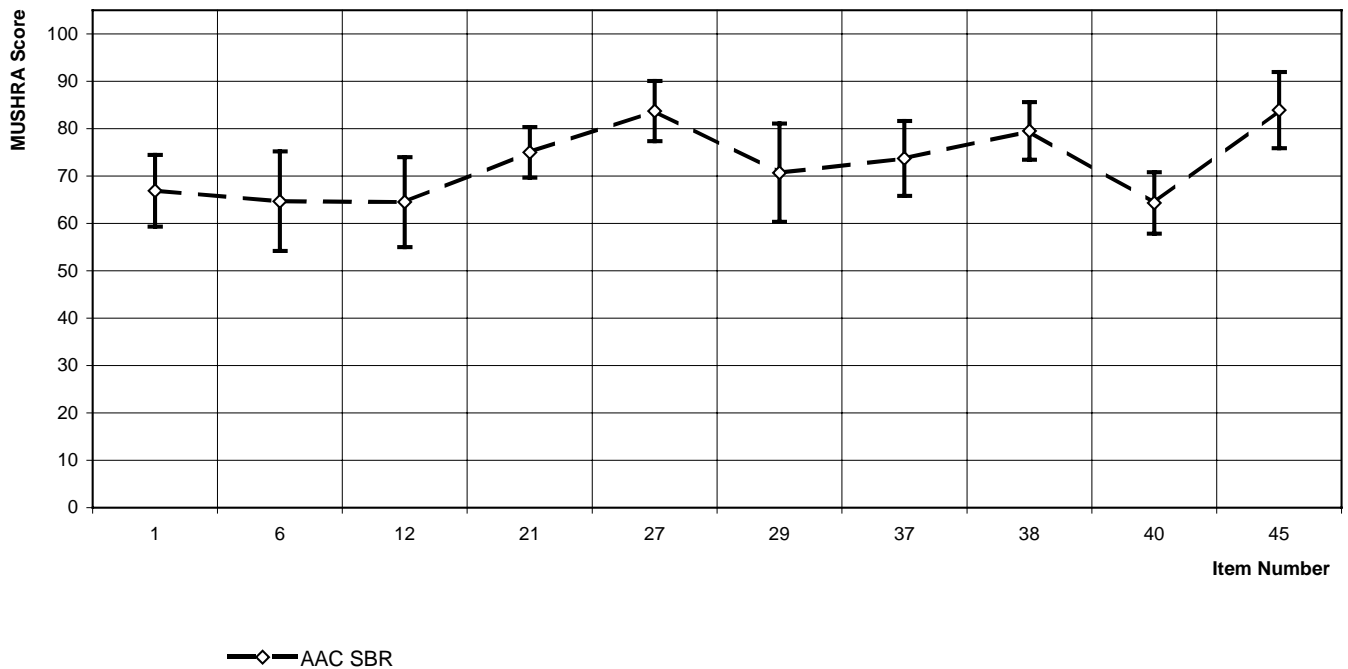




Graph 6 – Bosch Test Site: Means for All Items and Codecs

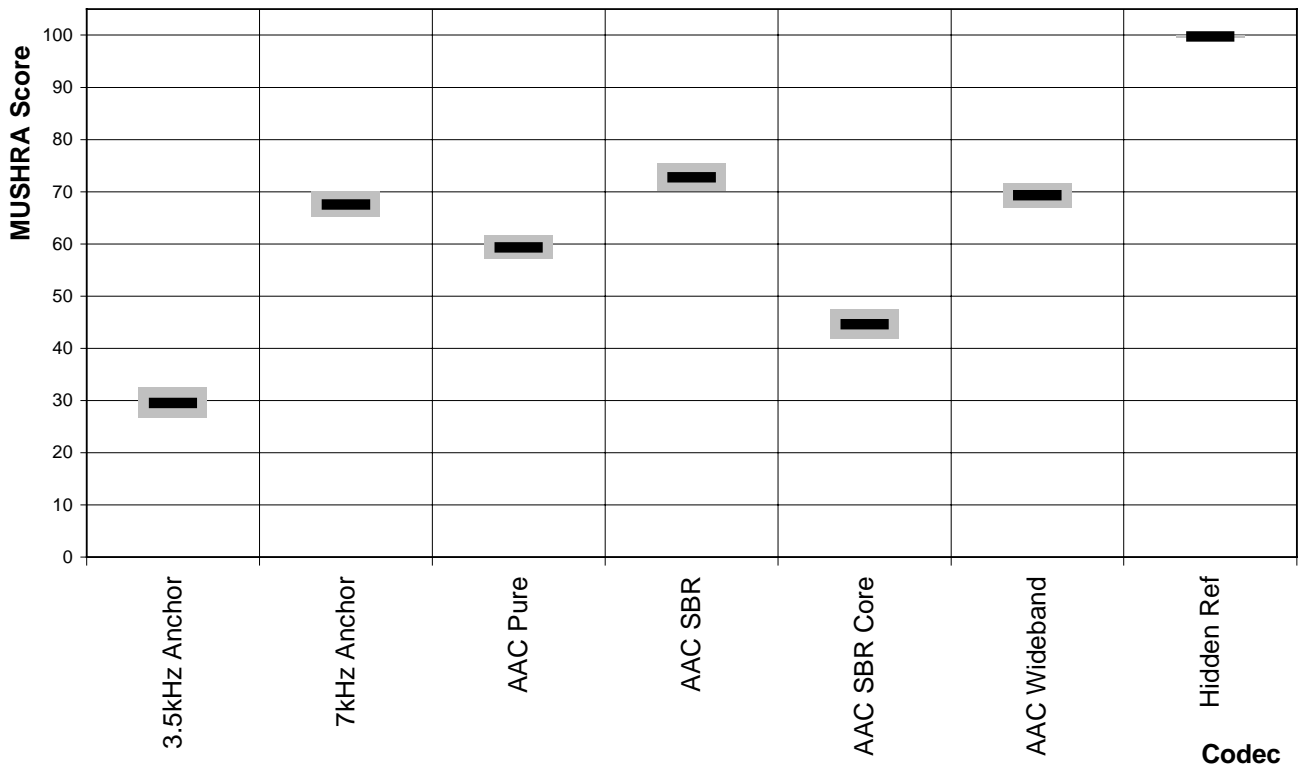


Graph 7 – Bosch Test Site: Mean and 95% Confidence Intervals for AAC SBR





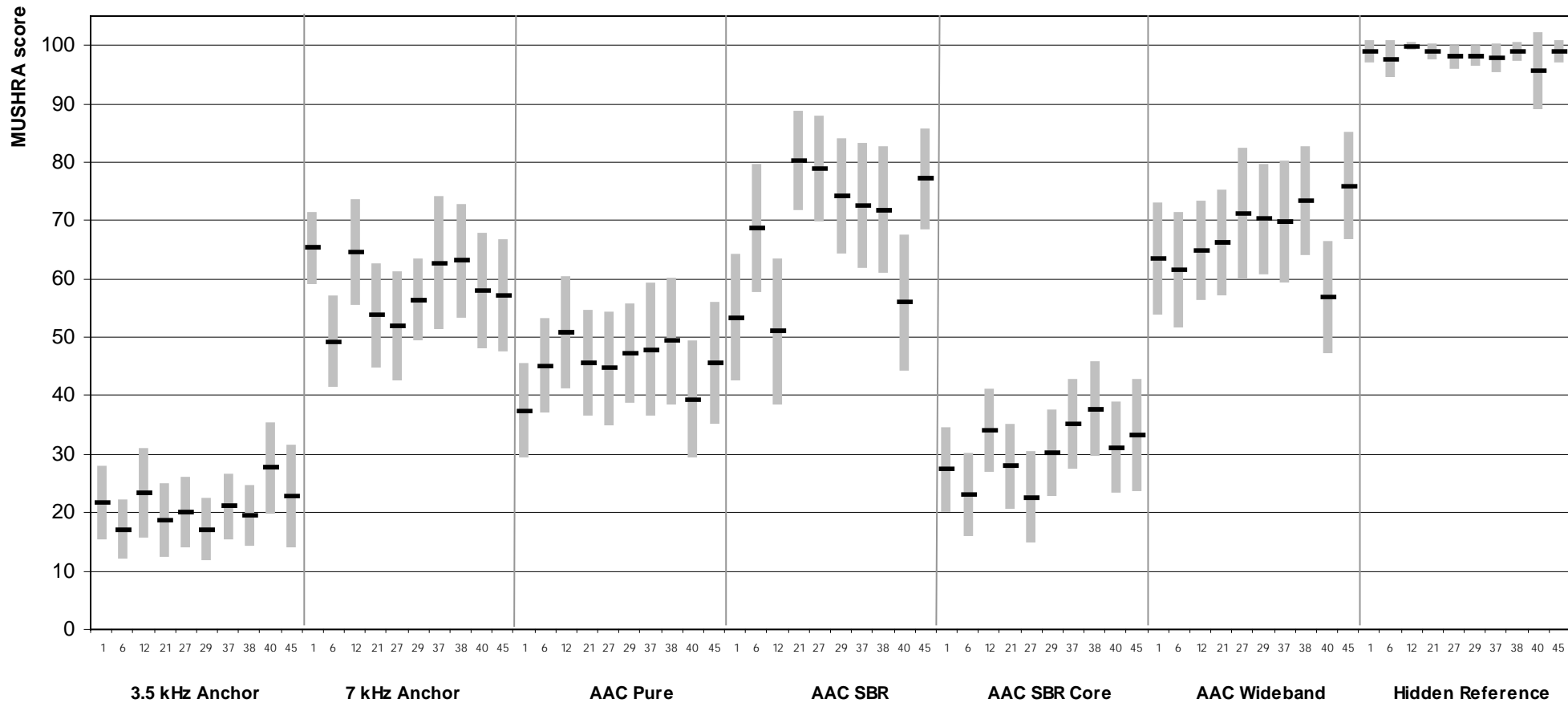
Graph 8 – Bosch Test Site: Mean and 95% Confidence Intervals for each Codec Over All Items





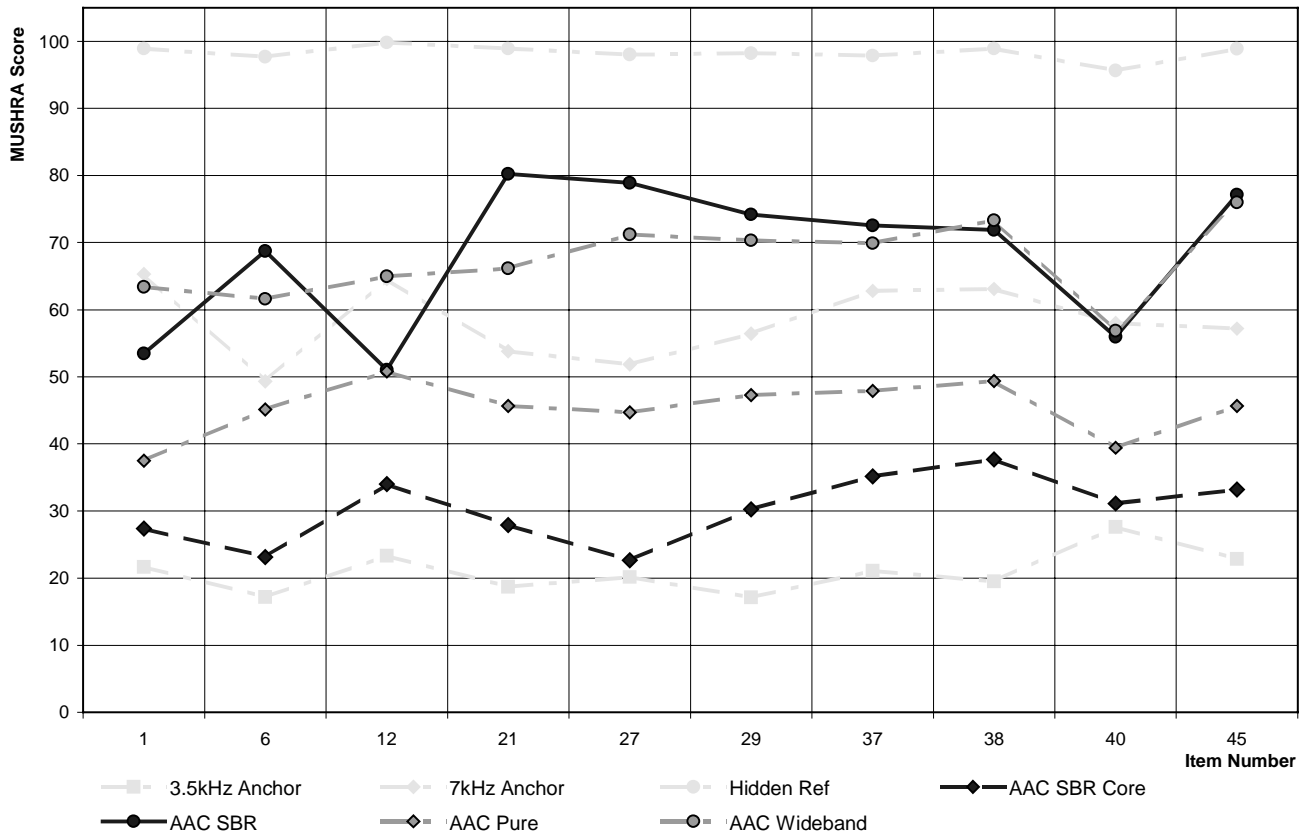
6.5 T-Nova Test Site

Graph 9 – T-Nova Test Site: Mean and 95% Confidence Intervals for All Items and Codecs

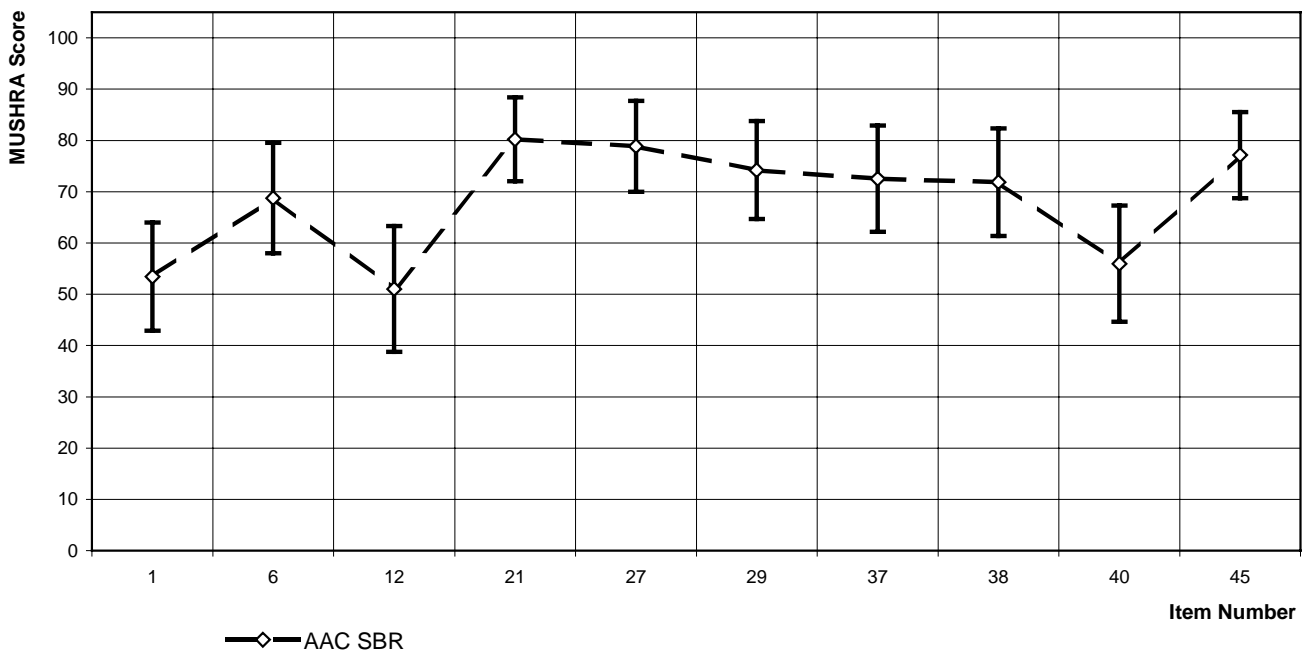




Graph 10 – T-Nova Test Site: Means for All Items and Codecs

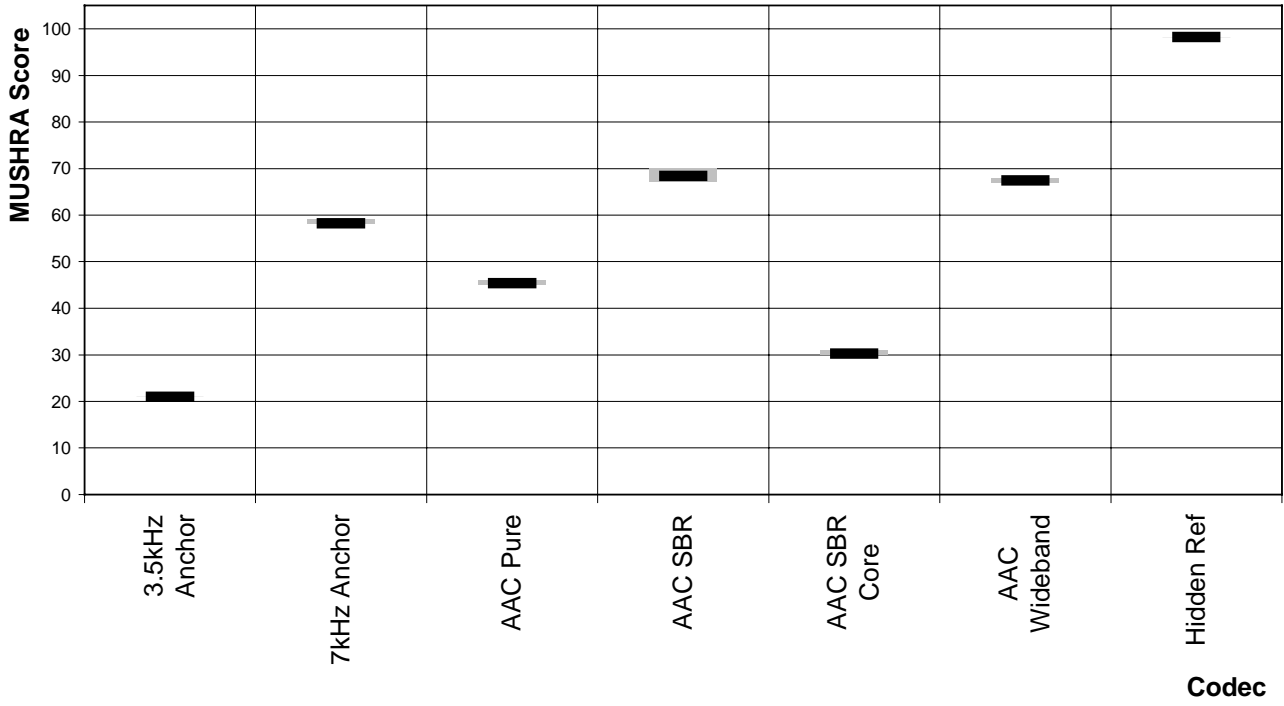


Graph 11 – T-Nova Test Site: Mean and 95% Confidence Intervals for AAC SBR





Graph 12 – T-Nova Test Site: Mean and 95% Confidence Intervals for each Codec Over All Items





7. Conclusions

1. Overall, AAC SBR performed significantly better than AAC Pure at all test sites. Looking item by item, at all test sites, AAC SBR averaged higher than AAC Pure. In addition, for a majority of items at BBC and T-Nova and 50% at Bosch, AAC SBR performed significantly better than AAC Pure. For some items (a few at BBC and T-Nova and 50% at Bosch) confidence intervals were overlapping.
2. Comparing AAC Wideband with AAC SBR, overall AAC SBR performed equally well (at Bosch and T-Nova) or better (BBC). Item per item, AAC SBR performed similarly (Bosch, T-Nova) or significantly better for 4 items and similarly for the others (BBC).

In summary, AAC SBR (24 kbit/s) performed at least as well as AAC Wideband (32 kbit/s). The quality achieved by AAC SBR in these tests at 24 kbit/s was therefore equal to the quality achieved with pure AAC at a 33% higher bit-rate (verified only for this bit rate).



References

1. C JTC1/SC29/WG11 (ref. MPEG98/N2425) , MPEG-4 Audio verification test results: Audio on Internet, October 1998, Eric Scheirer, Sang-Wook Kim, Martin Dietz
2. ISO/IEC JTC 1/SC 29/WG 11N3075 December 1999 Report on the MPEG-4 Audio Version 2 Verification Test, Ralph Sperschneider (FhG), Frank Feige (T-Nova), Schuyler Quackenbush (AT&T)
3. EBU Technical recommendation: MUSHRA-EBU Method for Subjective Listening Tests of Intermediate Audio Quality, Doc. B/AIM022, October 1999
4. ITU-R Recommendation BS.1116-1, “Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multi-channel Sound Systems”, 10/97



Annex A: Selection Panel Report

Report on the Selection Process for the Subjective Listening Tests of AAC-based DRM Audio Coding

Tasks assigned to the selection panel:

- select in total 10 critical items of typical broadcast material
- the material should include speech only, music only and speech together with (background) music
- find critical material for each of the coders (clearly audible impairments & different characteristics) to check if there is a difference (clearly audible impairments) between loudspeaker or headphone reproduction and report it
- select 3 or 4 suitable items used for training of the test persons

Listening panel from 11/13/00 – 11/14/00

Place Berlin at T-Nova Berkom

Members:

Andrew Murphy (BBC Research & Development)

Heiko Purnhagen (University Hanover)

Thomas Buchholz (T-Nova Deutsche Telekom)

Conclusions

Selection of the 10 critical items of typical broadcast material

The following 10 items of typical broadcast material were found to be critical for all of the codecs under test by the selection panel.

No.	Item number	Name	Category
1	Item_01	Male voice (English)	Speech only
2	Item_06	Folk music	Music only
3	Item_12	Speech + noise (Swedish)	Speech + Music
4	Item_21	Susan Vega, Tom's diner	Speech only
5	Item_27	Complex (sound+applause)	Music only
6	Item_29	"route 66"	Music only
7	Item_37	mainly speech (Spanish news) 1	Speech + Music
8	Item_38	mainly speech (Spanish news) 2	Speech + Music
9	Item_40	speech only (English feature)	Speech only
10	Item_45	music, speech different languages	Music only



Training Items

The following four items are recommended for training of the test subjects.

No.	Item number	Name	Category
1	Item_04	Speech + music	Speech + Music
2	Item_22	Speech (Male German speech)	Speech only
3	Item_23	Complex (Tracy Chapmann)	Music only
4	Item_26	Music (Palmtop boogie)	Music only

Selection process

All codec versions had been numbered from C1 to C6 by Mr. Schwalbe (T-Nova) to make sure that for none of the selection panel members the codec identification was known.

The selection panel listened to all of the 46 test signals with each coding system using headphones (STAX lambda). The method of presentation was first the original then the coded versions in various orders. If helpful the panel listened to the original again between Codec C3 and C4. Each member of the selection panel took notes about their impressions after having listened to each test signal.

After listening to all 46 test signals the panel discussed their impressions for each test signal and dropped the less impaired ones (see list of item reduction and item categories).

The remaining items were divided into the required three categories: Speech + Music, Speech only and Music only (see list of item reduction and item categories).

The selection panel continued by listening to all items for one codec version in one category using a single centre speaker. After having listened to the different items the panel discussed their impressions again and marked the different items either critical, nearly critical or not critical. If there were too many critical items within in one category, the selection panel listened to those showing the same kind of artefacts again to choose the most appropriate ones.

Finally the selection panel listened to the chosen ones again with headphones (STAX lambda) to ensure that listening to the samples with loudspeakers didn't affect the choices. No significant differences in audible artefacts were found between listening to the samples with headphones or loudspeakers and none of these differences would have affected our choices.

The selection of critical and training items was found with clear consensus amongst all members of the selection panel.



Remark:

Andreas Ehret (Coding Technologies) remarked that the references for item 09, item 43, item 44 and item 45 were probably already coded and proposed to drop these items. The members of the selection panel believe that this might be true but this situation could arise in a real-life broadcast transmission chain. These are therefore realistic and representative signals. Furthermore the selection panel thinks that the audio is still of suitable quality.

Listening room and technical equipment

The listening room (compliant to ITU-BS.1116) and test equipment were provided by T-Nova Deutsche Telekom, who will provide a description if requested. The selection panel felt that the equipment and situation were quite sufficient for the selection task.

The audio signal were stored on a PC having a digital soundcard (Creamware Pulsar) and routed through a digital NEXUS system to 24 bit D/A converters. The analogue signal was passed to the loudspeaker (Geithain electronic RL 900) through a level controller.



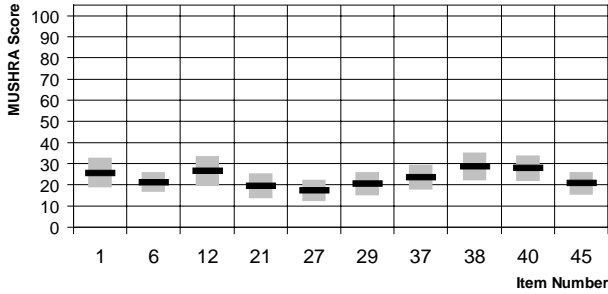
Item Reduction and Item Categories

No.	Description	Not Critical	Category
Item_01	Male voice (English)		speech only
Item_02	Female voice (English)		speech only
Item_03	Female voice (French)		speech only
Item_04	Speech + music		speech+music
Item_05	Pop music	x	
Item_06	Folk music		music only
Item_07	Male voice (French) + music		speech+music
Item_08	Music + noise		music only
Item_09	Female voice (singing)	x	
Item_10	Classic music (original 24 kHz)	x	
Item_11	Speech + noise	x	
Item_12	Speech + noise (Swedish)		speech+background
Item_13	Mozart: Requiem - beginning of Dies Irae		music only
Item_14	Female speech (Dutch) & Music		speech+music
Item_15	Female speech (Danish)		speech only
Item_16	Swedish Folk Music	x	
Item_17	Ice-hockey commentary		speech+background
Item_18	Lee Ritenour	x	
Item_19	Male speech (German)		speech+music
Item_20	Chris Rea - On the beach		music only
Item_21	Susan Vega, Tom's diner		speech only
Item_22	Speech		speech only
Item_23	Complex		music only
Item_24	Music	x	
Item_25	Music		music only
Item_26	Music		music only
Item_27	Complex		music+background
Item_28	Complex		music
Item_29	"route 66"		music only
Item_30	"route 66"		music only
Item_31	soprano solo	x	
Item_32	soprano solo		music only
Item_33	Abba		music only
Item_34	trumpet solo		music only
Item_35	mainly speech	x	
Item_36	mainly speech		speech only
Item_37	mainly speech		speech+music
Item_38	mainly speech		speech+music
Item_39	speech only	x	
Item_40	speech only		speech
Item_41	mainly speech		speech+music
Item_42	mainly speech	x	
Item_43	music, speech different languages		speech+music
Item_44	music, speech different languages	x	
Item_45	music, speech different languages		music only
Item_46	music, speech different languages	x	

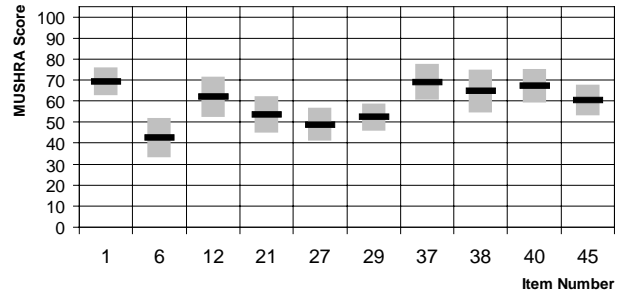


Annex B: Detailed Results BBC Test Site: Mean and 95% Confidence Intervals

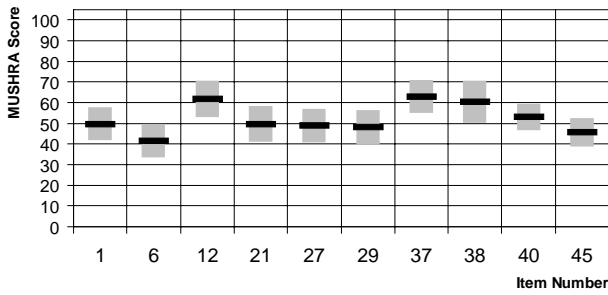
3.5 kHz Anchor



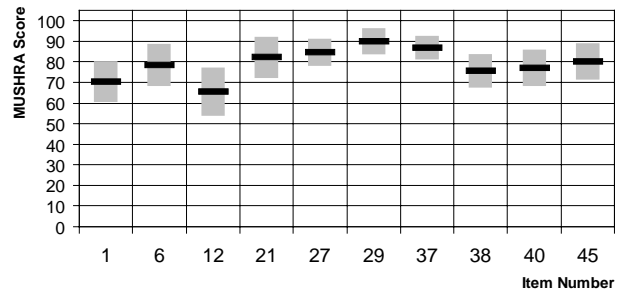
7 kHz Anchor



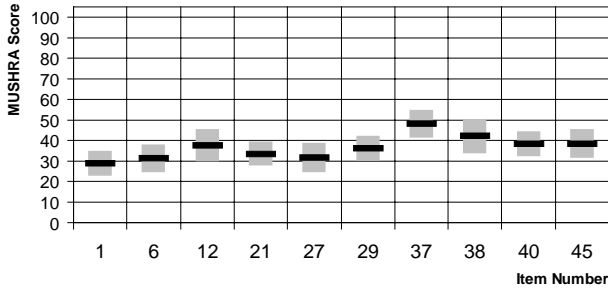
AAC Pure



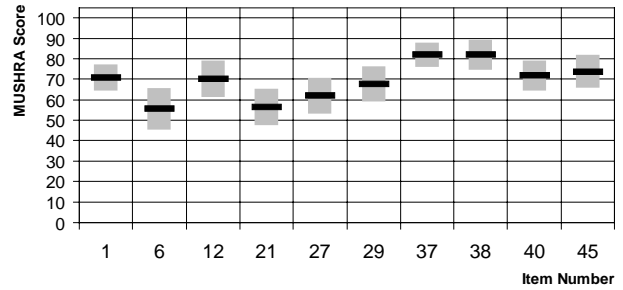
AAC SBR



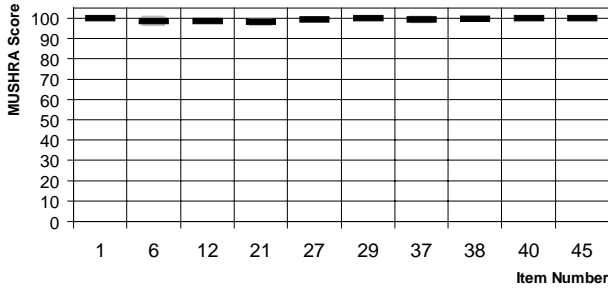
AAC SBR Core



AAC Wideband

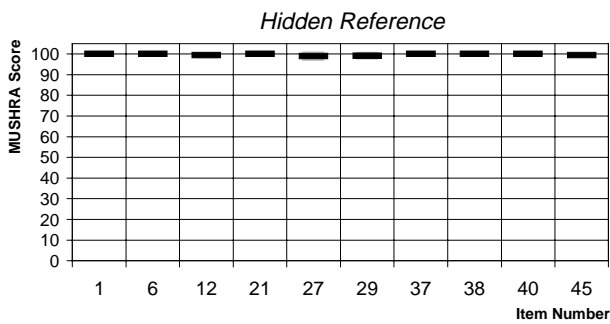
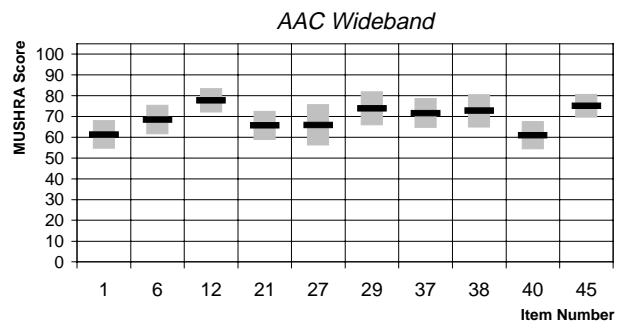
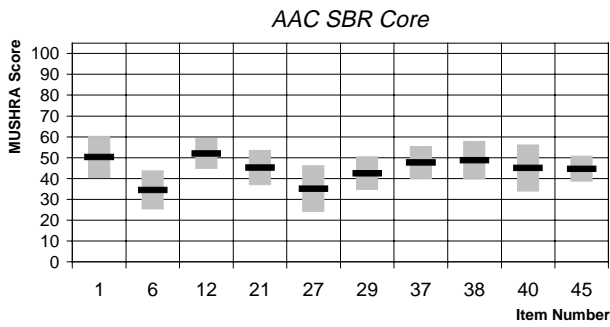
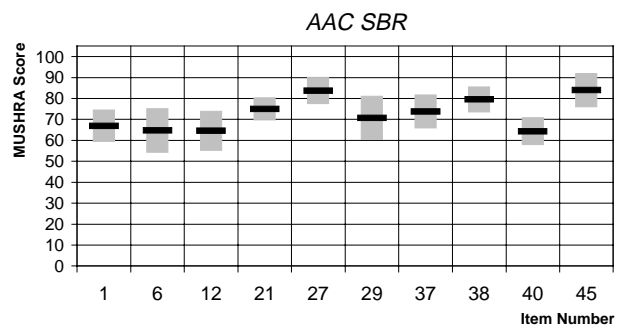
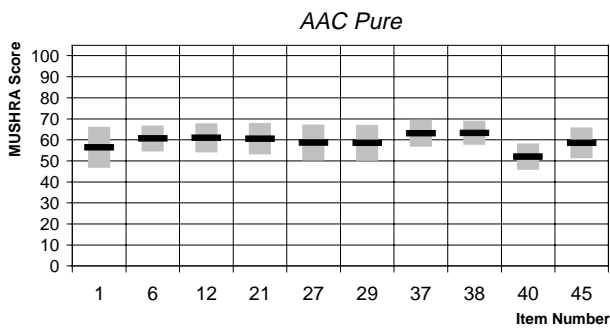
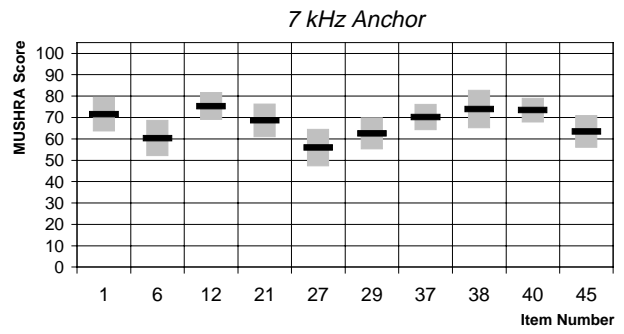
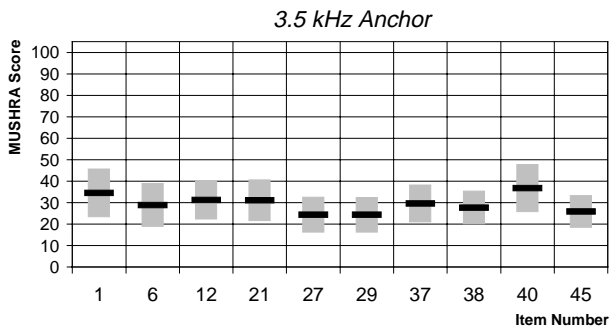


Hidden Reference



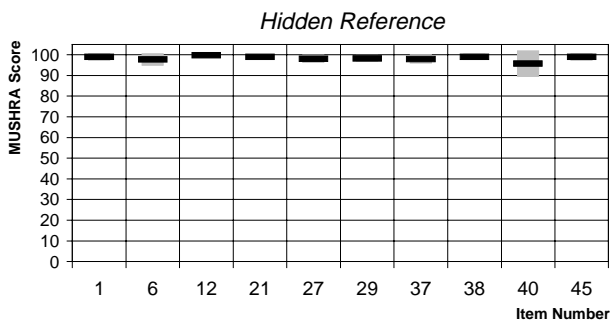
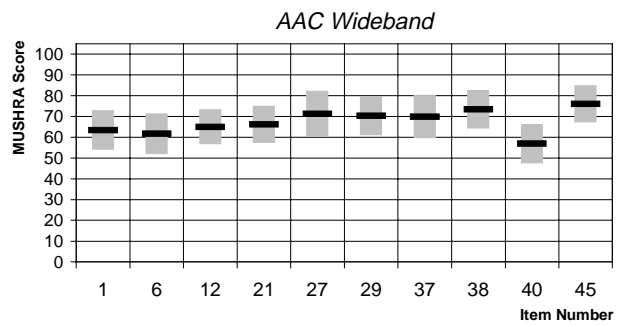
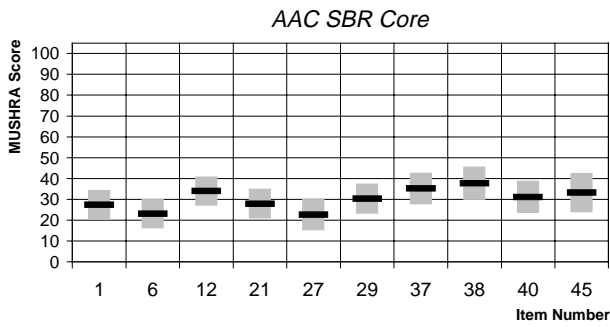
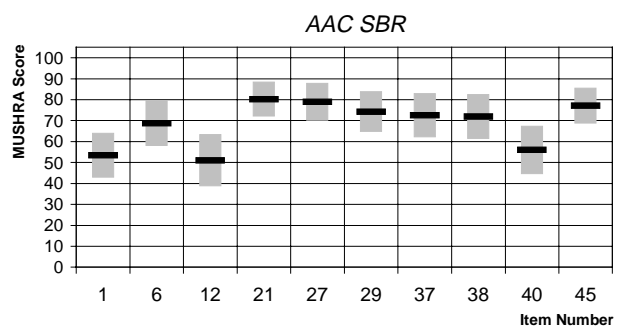
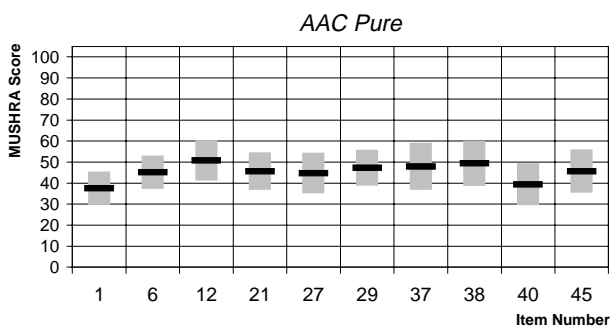
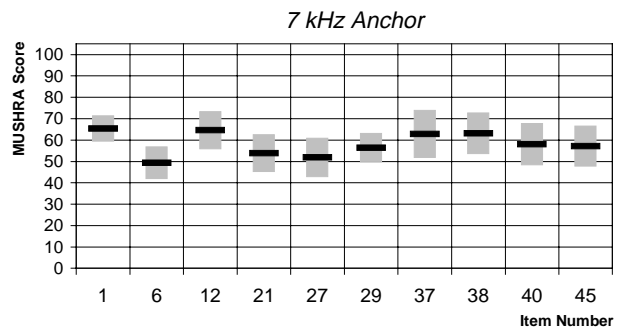
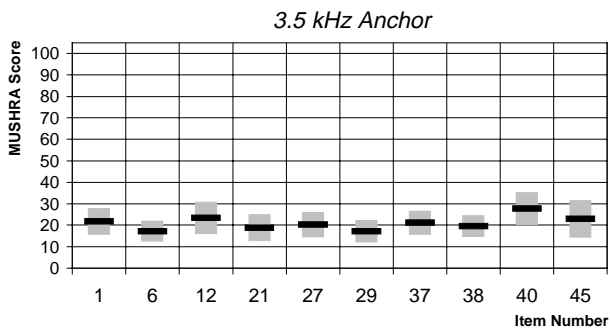


Annex B: Detailed Results Bosch Test Site: Mean and 95% Confidence Intervals





Annex B: Detailed Results T-Nova Test Site: Mean and 95% Confidence Intervals



Annex C: BBC Test site: Listening Room Conditions and Technical Equipment

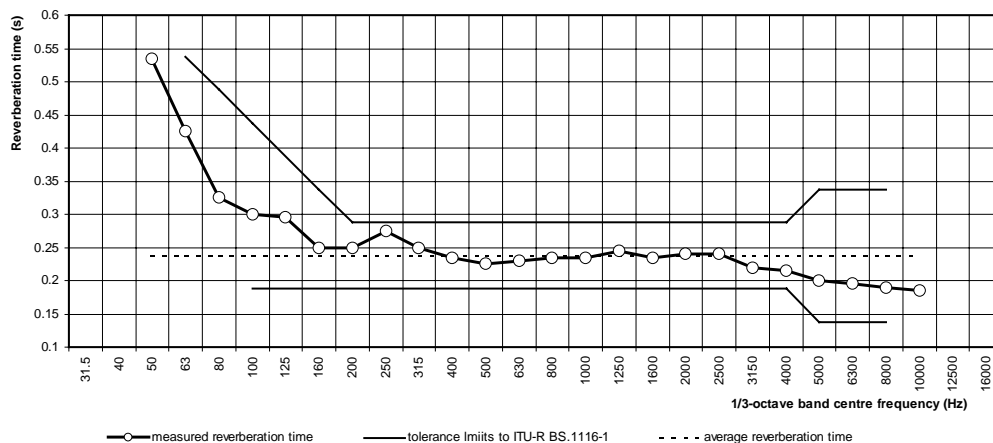
The tests took place in the reference listening room LR3 at BBC Research & Development, Kingswood Warren.

Room Dimensions and Layout

Length: 6.76 m
 Width: 4.93 m
 Height: 3.20 m
 Floor Area: 33.33 m²
 Room Volume: 106.6 m³

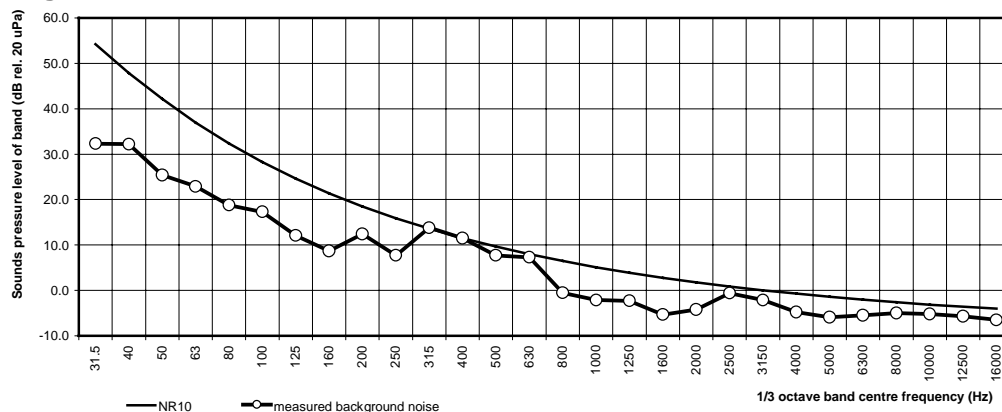
The speaker was situated symmetrically in the room, with its acoustic centre 1m from the back wall. The listener was situated 2m from the speaker, on axis.

Reverberation Time



The mean reverberation time between 200Hz and 2kHz is 0.238s which is in accordance with the ITU-R recommendations for this size of room.

Background Noise



The noise level at the reference listening position meets the noise criterion NR10.



Frequency Response Measurements

The operational room response curve was measured at the reference listening position in 1/3 octave bands using pink noise.



The operational room response meets the ITU-R required tolerances.

Equipment List

Device	Manufacturer	Description
PC	Dell	playout PC
Montego II	Turtle Beach	PC digital I/O card
Dream DA-1	Prism Sound	digital to analogue converter
		1/3 octave graphic equaliser
LS5/8	BBC	studio monitoring loudspeaker with Quad 405 power amplifier



Annex D: Tables of All Results BBC Test Site: Numerical Results

Results for 3.5 kHz low-pass filtered anchors

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	25.7	15.63	6.5
Item_06	folk music	21.3	10.19	4.3
Item_12	speech + noise (Swedish)	26.5	16.03	6.7
Item_21	Susan Vega. Tom's diner	19.5	13.04	5.4
Item_27	complex (sound+applause)	17.3	10.96	4.6
Item_29	"route 66"	20.4	12.48	5.2
Item_37	mainly speech (Spanish news) 1	23.4	13.09	5.5
Item_38	mainly speech (Spanish news) 2	28.6	15.06	6.3
Item_40	speech only (English feature)	27.8	13.40	5.6
Item_45	music, speech different languages	20.7	12.00	5.0

Results for 7 kHz low-pass filtered anchors

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	69.2	14.74	6.2
Item_06	folk music	42.5	21.27	8.9
Item_12	speech + noise (Swedish)	62.0	21.76	9.1
Item_21	Susan Vega. Tom's diner	53.5	19.53	8.2
Item_27	complex (sound+applause)	48.9	17.66	7.4
Item_29	"route 66"	52.4	14.36	6.0
Item_37	mainly speech (Spanish news) 1	69.0	19.20	8.0
Item_38	mainly speech (Spanish news) 2	64.8	23.49	9.8
Item_40	speech only (English feature)	67.2	17.78	7.4
Item_45	music, speech different languages	60.5	16.68	7.0

Results for AAC Pure

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	49.6	17.92	7.5
Item_06	folk music	41.4	18.21	7.6
Item_12	speech + noise (Swedish)	61.8	19.99	8.4
Item_21	Susan Vega. Tom's diner	49.6	19.79	8.3
Item_27	complex (sound+applause)	48.8	18.46	7.7
Item_29	"route 66"	48.0	18.87	7.9
Item_37	mainly speech (Spanish news) 1	63.0	18.04	7.5
Item_38	mainly speech (Spanish news) 2	60.2	23.39	9.8
Item_40	speech only (English feature)	53.0	14.34	6.0
Item_45	music, speech different languages	45.6	15.40	6.4



Results for AAC SBR

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	70.3	22.62	9.5
Item_06	folk music	78.5	23.33	9.8
Item_12	speech + noise (Swedish)	65.4	26.82	11.2
Item_21	Susan Vega. Tom's diner	82.1	23.06	9.6
Item_27	complex (sound+applause)	84.5	14.38	6.0
Item_29	"route 66"	89.9	14.10	5.9
Item_37	mainly speech (Spanish news) 1	86.8	12.27	5.1
Item_38	mainly speech (Spanish news) 2	75.7	18.37	7.7
Item_40	speech only (English feature)	76.9	20.04	8.4
Item_45	music, speech different languages	80.2	20.04	8.4

Results for AAC SBR Core

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	28.7	13.58	5.7
Item_06	folk music	31.2	15.00	6.3
Item_12	speech + noise (Swedish)	37.5	17.59	7.3
Item_21	Susan Vega. Tom's diner	33.5	13.06	5.5
Item_27	complex (sound+applause)	31.5	16.11	6.7
Item_29	"route 66"	36.2	13.73	5.7
Item_37	mainly speech (Spanish news) 1	48.2	15.00	6.3
Item_38	mainly speech (Spanish news) 2	42.0	18.86	7.9
Item_40	speech only (English feature)	38.3	13.13	5.5
Item_45	music, speech different languages	38.3	15.56	6.5

Results for AAC Wideband

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	70.9	14.47	6.0
Item_06	folk music	55.6	23.27	9.7
Item_12	speech + noise (Swedish)	70.1	20.18	8.4
Item_21	Susan Vega. Tom's diner	56.4	20.19	8.4
Item_27	complex (sound+applause)	61.9	19.70	8.2
Item_29	"route 66"	67.6	19.46	8.1
Item_37	mainly speech (Spanish news) 1	82.0	13.24	5.5
Item_38	mainly speech (Spanish news) 2	82.0	16.41	6.9
Item_40	speech only (English feature)	71.8	16.56	6.9
Item_45	music, speech different languages	73.7	18.24	7.6



Results for Hidden Reference

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	100.0	0.00	0.0
Item_06	folk music	98.7	5.15	2.2
Item_12	speech + noise (Swedish)	98.8	3.18	1.3
Item_21	Susan Vega. Tom's diner	98.2	4.46	1.9
Item_27	complex (sound+applause)	99.5	2.13	0.9
Item_29	"route 66"	100.0	0.00	0.0
Item_37	mainly speech (Spanish news) 1	99.3	3.20	1.3
Item_38	mainly speech (Spanish news) 2	99.7	1.49	0.6
Item_40	speech only (English feature)	100.0	0.00	0.0
Item_45	music, speech different languages	100.0	0.00	0.0



Annex D: Tables of All Results Bosch Test Site: Numerical Results

Results for 3.5 kHz low-pass filtered anchors

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	34.5	22.0	11.1
Item_06	folk music	28.9	19.9	10.1
Item_12	speech + noise (Swedish)	31.3	17.9	9.1
Item_21	Susan Vega. Tom's diner	31.1	19.0	9.6
Item_27	complex (sound+applause)	24.4	16.2	8.2
Item_29	"route 66"	24.3	16.2	8.2
Item_37	mainly speech (Spanish news) 1	29.6	17.1	8.7
Item_38	mainly speech (Spanish news) 2	27.7	15.3	7.7
Item_40	speech only (English feature)	36.7	21.8	11.0
Item_45	music, speech different languages	25.9	14.7	7.4

Results for 7 kHz low-pass filtered anchors

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	71.5	15.8	8.0
Item_06	folk music	60.3	16.4	8.3
Item_12	speech + noise (Swedish)	75.3	12.6	6.4
Item_21	Susan Vega. Tom's diner	68.6	15.2	7.7
Item_27	complex (sound+applause)	55.9	17.0	8.6
Item_29	"route 66"	62.5	14.6	7.4
Item_37	mainly speech (Spanish news) 1	70.2	11.8	6.0
Item_38	mainly speech (Spanish news) 2	73.9	17.5	8.9
Item_40	speech only (English feature)	73.4	11.1	5.6
Item_45	music, speech different languages	63.4	14.9	7.5

Results for AAC Pure

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	56.4	18.9	9.6
Item_06	folk music	60.6	11.9	6.0
Item_12	speech + noise (Swedish)	60.9	13.3	6.7
Item_21	Susan Vega. Tom's diner	60.5	14.4	7.3
Item_27	complex (sound+applause)	58.7	16.5	8.4
Item_29	"route 66"	58.5	16.6	8.4
Item_37	mainly speech (Spanish news) 1	63.1	12.3	6.2
Item_38	mainly speech (Spanish news) 2	63.3	11.0	5.6
Item_40	speech only (English feature)	52.0	12.0	6.1
Item_45	music, speech different languages	58.5	14.1	7.1



Results for AAC SBR

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	66.9	14.9	7.5
Item_06	folk music	64.7	20.7	10.5
Item_12	speech + noise (Swedish)	64.5	18.7	9.5
Item_21	Susan Vega. Tom's diner	75.0	10.6	5.4
Item_27	complex (sound+applause)	83.7	12.5	6.3
Item_29	"route 66"	70.7	20.5	10.4
Item_37	mainly speech (Spanish news) 1	73.7	15.6	7.9
Item_38	mainly speech (Spanish news) 2	79.5	12.0	6.1
Item_40	speech only (English feature)	64.3	12.8	6.5
Item_45	music, speech different languages	83.9	15.9	8.0

Results for AAC SBR Core

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	50.3	19.6	9.9
Item_06	folk music	34.5	18.1	9.2
Item_12	speech + noise (Swedish)	52.0	14.3	7.2
Item_21	Susan Vega. Tom's diner	45.3	16.4	8.3
Item_27	complex (sound+applause)	35.1	21.8	11.0
Item_29	"route 66"	42.5	15.6	7.9
Item_37	mainly speech (Spanish news) 1	47.6	15.3	7.7
Item_38	mainly speech (Spanish news) 2	48.7	17.9	9.1
Item_40	speech only (English feature)	45.0	22.1	11.2
Item_45	music, speech different languages	44.7	12.1	6.1

Results for AAC Wideband

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	61.3	13.2	6.7
Item_06	folk music	68.4	13.7	6.9
Item_12	speech + noise (Swedish)	77.7	11.4	5.8
Item_21	Susan Vega. Tom's diner	65.6	13.5	6.8
Item_27	complex (sound+applause)	65.9	19.3	9.8
Item_29	"route 66"	73.9	15.8	8.0
Item_37	mainly speech (Spanish news) 1	71.6	13.8	7.0
Item_38	mainly speech (Spanish news) 2	72.7	15.5	7.8
Item_40	speech only (English feature)	60.9	13.0	6.6
Item_45	music, speech different languages	75.1	11.1	5.6



Results for Hidden Reference

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	100	0.0	0.0
Item_06	folk music	100	0.0	0.0
Item_12	speech + noise (Swedish)	99	2.6	1.3
Item_21	Susan Vega. Tom's diner	100	0.0	0.0
Item_27	complex (sound+applause)	99	4.1	2.1
Item_29	"route 66"	99	3.4	1.7
Item_37	mainly speech (Spanish news) 1	100	0.0	0.0
Item_38	mainly speech (Spanish news) 2	100	0.0	0.0
Item_40	speech only (English feature)	100	0.0	0.0
Item_45	music, speech different languages	99	1.6	0.8



Annex D: Tables of All Results T-Nova Test Site: Numerical Results

Results for 3.5 kHz low-pass filtered anchors

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	21.7	15.10	6.0
Item_06	folk music	17.2	11.82	4.7
Item_12	speech + noise (Swedish)	23.3	18.52	7.4
Item_21	Susan Vega. Tom's diner	18.7	15.36	6.1
Item_27	complex (sound+applause)	20.2	14.42	5.8
Item_29	"route 66"	17.1	12.65	5.1
Item_37	mainly speech (Spanish news) 1	21.1	13.50	5.4
Item_38	mainly speech (Spanish news) 2	19.5	12.26	4.9
Item_40	speech only (English feature)	27.6	18.94	7.6
Item_45	music, speech different languages	22.9	21.34	8.5

Results for 7 kHz low-pass filtered anchors

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	65.3	14.90	6.0
Item_06	folk music	49.3	18.83	7.5
Item_12	speech + noise (Swedish)	64.6	21.96	8.8
Item_21	Susan Vega. Tom's diner	53.8	21.56	8.6
Item_27	complex (sound+applause)	51.9	22.59	9.0
Item_29	"route 66"	56.4	16.90	6.8
Item_37	mainly speech (Spanish news) 1	62.8	27.69	11.1
Item_38	mainly speech (Spanish news) 2	63.1	23.96	9.6
Item_40	speech only (English feature)	58.0	24.36	9.7
Item_45	music, speech different languages	57.2	23.58	9.4

Results for AAC Pure

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	37.5	19.49	7.8
Item_06	folk music	45.1	19.57	7.8
Item_12	speech + noise (Swedish)	50.8	23.52	9.4
Item_21	Susan Vega. Tom's diner	45.7	21.97	8.8
Item_27	complex (sound+applause)	44.7	23.81	9.5
Item_29	"route 66"	47.3	20.89	8.4
Item_37	mainly speech (Spanish news) 1	47.9	27.64	11.1
Item_38	mainly speech (Spanish news) 2	49.4	26.39	10.6
Item_40	speech only (English feature)	39.4	24.51	9.8
Item_45	music, speech different languages	45.7	25.22	10.1



Results for AAC SBR

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	53.5	26.34	10.5
Item_06	folk music	68.8	26.89	10.8
Item_12	speech + noise (Swedish)	51.0	30.74	12.3
Item_21	Susan Vega. Tom's diner	80.2	20.44	8.2
Item_27	complex (sound+applause)	78.9	22.13	8.9
Item_29	"route 66"	74.2	23.88	9.6
Item_37	mainly speech (Spanish news) 1	72.5	25.87	10.4
Item_38	mainly speech (Spanish news) 2	71.9	26.23	10.5
Item_40	speech only (English feature)	56.0	28.40	11.4
Item_45	music, speech different languages	77.1	21.04	8.4

Results for AAC SBR Core

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	27.4	17.43	7.0
Item_06	folk music	23.2	17.24	6.9
Item_12	speech + noise (Swedish)	34.0	17.06	6.8
Item_21	Susan Vega. Tom's diner	27.9	17.56	7.0
Item_27	complex (sound+applause)	22.7	18.71	7.5
Item_29	"route 66"	30.3	17.62	7.1
Item_37	mainly speech (Spanish news) 1	35.2	18.50	7.4
Item_38	mainly speech (Spanish news) 2	37.7	19.49	7.8
Item_40	speech only (English feature)	31.1	18.90	7.6
Item_45	music, speech different languages	33.2	23.35	9.3

Results for AAC Wideband

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	63.4	23.36	9.3
Item_06	folk music	61.6	24.13	9.7
Item_12	speech + noise (Swedish)	65.0	20.63	8.3
Item_21	Susan Vega. Tom's diner	66.2	21.90	8.8
Item_27	complex (sound+applause)	71.2	27.34	10.9
Item_29	"route 66"	70.3	22.99	9.2
Item_37	mainly speech (Spanish news) 1	69.9	25.47	10.2
Item_38	mainly speech (Spanish news) 2	73.3	22.66	9.1
Item_40	speech only (English feature)	56.8	23.42	9.4
Item_45	music, speech different languages	76.0	22.08	8.8



Results for Hidden Reference

Item No	Name	Mean score	Standard Deviation	95% Confidence Interval
Item_01	male voice (English)	98.9	4.12	1.6
Item_06	folk music	97.7	7.12	2.9
Item_12	speech + noise (Swedish)	99.8	1.02	0.4
Item_21	Susan Vega. Tom's diner	98.9	2.52	1.0
Item_27	complex (sound+applause)	98.0	4.73	1.9
Item_29	"route 66"	98.2	3.73	1.5
Item_37	mainly speech (Spanish news) 1	97.9	5.23	2.1
Item_38	mainly speech (Spanish news) 2	98.9	3.48	1.4
Item_40	speech only (English feature)	95.6	15.69	6.3
Item_45	music, speech different languages	98.9	4.14	1.7