EBU – TECH 3324

EBU Evaluations of Multichannel Audio Codecs

Status: Report

Source: D/MAE

Geneva
September 2007

Page intentionally left blank. This document is paginated for recto-verso printing

Contents

EBU evaluations of multichannel audio codecs

EBU Committee	*First Issued*	*Revised*	*Re-issued*
DMC	2007		

Keywords: Multichannel Audio, Codec, MUSHRA

1. Introduction

With the advent of High Definition television services (HDTV), the public is increasingly being exposed to surround sound presentations using so-called home theatre environments. Multichannel audio is available also from recorded media and computer games. However, the restricted bandwidth available into the home, whether by broadcast (satellite, terrestrial, or cable), or via broadband, means that there is an increasing interest in the performance of low bit rate surround sound audio coding systems for "emission" coding.

The European Broadcasting Union Project Group B/MAE (Multichannel Audio Evaluations) was set up in 2006 in order to perform the assessment of the sound quality of some multichannel audio bit rate reduction codecs for broadcast applications. Codecs under test include offerings from Dolby, DTS, Microsoft, implementations of MPEG AAC and of the new MPEG Surround codec. The bit rates ranged from 64 kbit/s to 1.5 Mbit/s.

The last time that the EBU conducted subjective quality tests on surround sound codecs there were few available [1]. Now the range is much wider, little test data is publicly available and so the choice is more difficult. To help broadcasters make informed decisions, an EBU technical project group has performed extensive subjective tests of a wide range of codecs at a wide range of bit rates on a wide range of test material. The test method and the results are described in this report.

The choice of test method was made based on the expected range of quality. Of the two main methods, BS.1116 [2] and MUSHRA [3], it was felt that MUSHRA was more appropriate, at least for initial tests. If finer discrimination between high quality systems were to be required, then BS.1116 tests could have been conducted on a subset of the systems later. However MUSHRA was found to provide satisfactory resolution of all the test results, therefore BS. 1116 was not used at all.

The choice of test material followed the process specified in the test method recommendations. The call for proposed items brought in a wide variety of new and interesting material, showing the advances that have been made in surround sound production since the first EBU tests. Ten items were selected for testing, and four for training of subjects.

The testing was conducted by six test laboratories, each receiving a random selection of the codec and bit rate combinations. Each laboratory aimed to have at least 15 subjects listen to each test stimulus. This resulted in over 14000 measurements.

The statistical analysis focussed on mean opinion scores and 95% confidence intervals. A variety of methods for detection and eliminating unreliable subjects' data were also investigated.

During the B/MAE preparatory phase some difficulties were encountered. One codec proponent (Coding Technologies) discovered that its codec (HE AAC) had a severe bug, however this discovery was made only after the test sequences have been revealed to the proponents and after the codec submission deadline. This codec had to be withdrawn from the Phase 1 tests and included it in Phase 2, along with some other new codecs. The two Phases used identical evaluation methodologies, however the test sequences selected were different.

There is further work to be done. The cascaded stereo codec tests [4] conducted by the same EBU technical project group, in which a broadcast chain of five concatenated codecs was used, could well now be repeated with surround sound codecs. Unfortunately, this represents an enormous amount of effort, so in the short term, some tests on transcoding will be done. The scope of these is to assess the impact of using a low bit rate emission coder for delivery to the home, followed by transcoding to a higher bit rate bit stream for distribution in the home over an S/PDIF connection for example as DTS at 1.5 Mbit/s or Dolby Digital at 448 kbit/s.

2. Participating Test Sites

As the number of test vectors in these tests was enormous, a single laboratory would not be able to perform all the tests required or else the tests would take too much time. Therefore, the work was divided among several EBU laboratories. Each laboratory performed part of the overall test workload, such that each test vector was performed by at least two laboratories. In this way it was possible to assess whether or not results from different laboratories are sufficiently well correlated. The MUSHRA methodology used in these tests is suited to allow for sharing the overall workload among several laboratories. The work was evenly divided among the following eight EBU laboratories:

- BBC (British Broadcasting Corporation)
- IRT (Institut für Rundfunktechnik, Munich, Germany)
- TVP (Polish Television, Warsaw, Poland)
- FT Orange (France Télécom, Rennes, France)
- RF (Radio France, Paris)
- RAI CRIT (RAI Research Centre, Turin, Italy)
- BR (Bayerischer Rundfunk, Munich, Germany)
- ORF (Austrian Broadcasting Organisation, Vienna, Austria)

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

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

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

3. Selected Codecs for Testing

These EBU tests involved several multichannel audio codecs fulfilling one of the two following criteria:

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

In order to identify the codecs corresponding to the above conditions, Project Group B/MAE analysed the whole broadcast chain including production, contribution, distribution and emission. Multichannel audio codecs are being used in radio, television and internet. It is likely however that different codecs (with different parameters) will be used in the studio, contribution, distribution, emission and in the home environment.

The tables below review possible candidate MCA codecs for a range of applications in the reference broadcast chain.

PRODUCTION AND ARCHIVING	
APPLICATION	**CANDIDATE MCA CODEC**
IT-based (RF64 Files)	DD DTS Dolby-E MPEG Layer II with 128 kbit/s per channel
IT-based (MXF wrapper)	DD DTS Dolby-E MPEG Layer II with 128 kbit/s per channel

CONTRIBUTION from OB van to broadcast house	
APPLICATION	**CANDIDATE MCA CODEC**
IP UDP/RTP	AAC apt-X 16/24 bit enhanced apt-X 16/24 bit linear PCM DTS DD
ISDN	AAC apt-X 16/24 bit enhanced apt-X 16/24 bit
T1/E1 for radio	emission codec (i.e. no additional coding) Dolby E DTS apt-X 16/24 bit (576 for stereo 24 bit and 3 times 384 (16bit) for MCA) enhanced apt-X 16/24 bit J.41 J.57 MPEG-1 Layer II at 192 kbit/s per channel
ASI and SDI embedded for TV	emission codec Dolby E DTS
AES/EBU with compressed MCA	Dolby E emission codec DTS
MADI	Uncompressed

DISTRIBUTION from broadcast house to playout	
APPLICATION	**CANDIDATE MCA CODEC**
T1/E1-lines for Radio (ASI over E1-lines)	emission codec, i.e. no additional encoding Dolby-E DTS
ASI and SDI-embedded for TV	emission codec Dolby-E DTS
MPLS	emission codec Dolby-E DTS
AES/EBU with compressed MCA	Dolby-E J.57

EMISSION from Transmitter to Receiver	
APPLICATION	**CANDIDATE MCA CODEC**
SDTV DVB-S/C/T	DD DD+ DTS AAC HE AAC (with MPEG Surround) Layer II MPEG Surround
Radio on DVB-S	DD DTS AAC HE AAC (with MPEG Surround for low bit rates) Layer II MPEG Surround
IPTV: SDTV and HDTV	DD DD+ DTS AAC HE AAC (with MPEG Surround for low bit rates) Layer II MPEG Surround
HDTV DVB-S/C/T	DD DD+ DTS AAC HE AAC
IP radio (IP multicast)	AAC HE AAC HE AAC MPEG Surround Windows Media MP-3 Surround Ogg Vorbis DTS DD+
Internet streaming (IP Unicast)	AAC HE AAC HE AAC MPEG Surround Windows Media MP-3 Surround Ogg Vorbis DTS DD+

DAB	Layer II MPEG Surround HE AAC HE AAC MPEG Surround
Broadcast to Handhelds: DVB-H	HE AAC HE AAC MPEG Surround
Broadcast to Handhelds: DMB	HE AAC HE AAC MPEG Surround

The present EBU tests focused on emission codecs only. As explained above due to a bug discovered in one of the codecs, the B/MAE Group decided to split the process into two parts, Phase 1 and Phase 2.

The following sections list the emission codecs evaluated in Phase 1 and Phase 2, respectively.

3.1 Phase 1

The following emission codecs were used in Phase 1:

Key	Codec	Bitrate
A	L2 MPS 224	224
B	WMA9 448	448
E	DD+ 200	200
G	WMA10 192	192
I	DD+ 224	224
K	AAC 320	320
L	DD+ 448	448
M	AAC 256	256
N	Spatial Anchor	X
O	HE AAC MPS 64	64
OR	Original	X
P	WMA9 192	192
Q	DTS 1,5	1500
R	WMA9 256	256
S	DD+ 256	256
T	DTS 448	448
U	DD 384	384
V	HE AAC MPS 96	96
W	Prologic2 L2 256	256
X	MP3 S 192	192
Y	Low Anchor	X
Z	WMA10 256	256

AAC: Advanced Audio Coding
DD: Dolby Digital (also known as AC3)
DD+ : Dolby Digital Plus
DTS: Digital Theatre System
HE AAC MPS: High Efficiency AAC MPEG Surround
L2 MPS: MPEG I Layer II MPEG Surround
MP3 S: MPEG I Layer III Surround
Prologic 2 L2: Dolby Prologic 2 MPEG I Layer II
WMA9: Windows Media Audio Version 9
WMA10: Windows Media Audio Version 10

3.2 Phase 2

The following emission codecs were used in Phase 2:

Key	Codec	Bitrate
A	DD+ 200	200
B	DD+ 256	256
C	HE AAC 128	128
D	HE AAC 160	160
E	HE AAC 192	192
F	L2 MPS 256	256
G	DD+ 200 ev	200
H	DD+ 256 ev	256
I	AAC 320	320
J	DD+ 256&DD+ 640	256&640
K	HE AAC 192&DTS 1500	192&1500
L	AAC 320&DTS 1500	320&1500
M	DD 448	448
O	Original	X
P	Spatial Anchor	X
Q	Low Anchor	X

DD: Dolby Digital (also known as AC3)
DD+ : Dolby Digital plus (new version)
DD+ ev: Dolby Digital plus (earlier version, under test in Phase 1)
HE AAC: High Efficiency AAC (new version, earlier version was redrawn by the manufacturer during Phase 1)
AAC xxx&DTS yyy: concatenation of AAC followed by DTS
DD+ xxx&DD+ yyy: concatenation of DD+ followed by DD+

4. Codec Parameters

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

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

Input resolution: at least 16 bit, possibly 24 bits

Encoding parameters: pre-configured by the codec developers

Alignment of test items (Input level to encoder): PML (Permitted Maximum Level): -9 dB FS measured with PPM (Peak Programme Meter) with 10 ms integration time (EBU Rec. 645). Clipping of the signal was avoided.

The above parameters did not change during the tests.

5. Test Sequences

5.1 Phase 1

Test sequences were audio excerpts chosen from typical radio and television programme services.

Nominal length of test sequences: about 15 seconds

Format: Unprocessed PCM (WAV or AIFF) and DSD (SACD). Dolby-E and processed materials using Broadcast processors (such as Optimod) were possible

Pre-selection process: Formal tests were preceded by a pre-selection process, during which the number of test sequences was reduced to ten. In addition, the pre-selection panel selected four items to be used for training. The two tables below give a list of sequences.

A copy of the full pre-selection report is given in **Appendix 1**.

Table 1: Items selected for test sessions

No.	Name	Description
1	Applause	applause with distinct clapping sounds
2	Sax_Piano	saxophone and piano
3	R_Plant_Rock	rock music (“Whole lotta love”)
4	Tschaikowsky_04	orchestral piece; open sound
5	Moonriver_Mancini	mouth organ and string orchestra
6	Sedambonjou_Salsa	atmospheric performance of Latin-American music
7	Harpsichord*	solo harpsichord; isolated notes
8	svt_Female-vocal	live performance of female vocalist and band; crowd sounds
9	hbv_Gregorian-chant	small choir; large church; Gregorian chant
10	Bach_organ2	church organ; lots of stops out

Table 2: Items selected for training sessions (4)

No.	Name	Description
1	Hadouk	Eastern woodwind, strings, percussion
2	Exodus	orchestral; lots of brass instruments
3	Hoelsky	“chattering” choral voices
4	Bach_organ1	church organ; few stops out

5.2 Phase 2

The following test items were used in Phase 2:

Table 3: Phase 2 test items

No.	Name	Description
1	Radio drama (Hörspiel) (tenorRP)	Clarinet, orchestra, male speaker, tenor, ambience
2	Harpsichord* (harpsic)	Bach: solo from harpsichord concert; isolated notes
3	Mouth organ (mouthha)	mouth organ, bass guitar, percussion
4	Trumpet (trumpet)	orchestral piece;
5	Orchestra (Lohengrin) (hornWag)	orchestral piece; string orchestra
6	Applause (from Phase 1) (applaus)	Applause with distinct clapping sounds

7	Exodus (from Phase 1) (brassEX)	orchestral; lots of brass instruments
8	transient guitar (fleetwd)	male vocalist and guitar
9	Choir (mtChoir)	choir; large church; British traditional
10	Organ (bach565)	Bach d-minor toccata; church organ; lots of stops out

Table 4: Phase 2 items selected for training session

No.	Name	Description
1	Jazz	Jazz Burghausen; live performance with applause
2	Violin	Vivaldi –Violin concert; solo violin and orchestra;
3	de Falla	Orchestra with trumpet, trombone, bass drum, snappers
4	Chris Botti Trott	saxophone, percussions, bass

*The Harpsichord item in Phase 1 was a different piece from that used in Phase 2 of the same name.

6. Encoding Process

6.1 Codecs

In order to evaluate subjectively the performance of codecs, all test sequences need to be encoded (and decoded) using the selected codecs at the chosen bit rates. Encoding of the test sequences was performed by IRT. In order to simplify the pre-selection process, only the minimum and maximum bit-rate were encoded if more than two bit rates per codec were used. For encoding and decoding of the ten test items used for the evaluation process (as well as the four training items), each chosen bit-rate was used.

The input signal required for encoding consisted of six individual audio channels, i.e. 6 mono wav-files. The DTS encoder was an exception - it required three stereo wav-files, i.e. L/R, C/LFE and Ls/Rs as input format. These three stereo pairs were produced by the Steinberg Nuendo audio editing program. The encoders produced the encoded files, which were used to verify the actual bit rate.

The time for the encoding process depended on the encoding scheme. All codecs, except the Dolby AC3 were software codecs, working in non real-time. The slowest encoding/decoding process was observed at the MPEG Surround codecs. The encoding, as well as the decoding time took about five times the duration of test sequence. Other codecs took about 0.2 to 0.1 of the real-time.

Only one codec, the Dolby AC3, performed encoding in real-time. Three AES pairs were fed into the input of the encoder. The resulting Dolby Digital bitstream was fed directly to the Dolby Digital decoder, which produced again three stereo AES pairs.

All software codecs had been installed on a Windows XP PC from FUJITSU-SIEMENS with 2.4 GHz CPU speed and 512 MByte internal memory.

The encoding and decoding processes in Phase 1 and Phase 2 were identical, however some other new codecs, or a new versions of the same codec, and somewhat different bit-rates were used. To this end, Phase 2 required some additional encoding effort.

The following table shows the version number of the various codecs used to generate the test and training sequences in Phase 1:

Table 5: Codec Versions (Phase 1)

Codecs - Phase 1	Version
Dolby Digital	Encoder Model DP 569, Firmware version: v.2.0.3.1 Decoder Model DP 562
Dolby Digital plus	Dolby Digital plus Prototype Encoder Version 1.4.0.2 Dolby Digital plus Decoder-Converter Simulation Version 1.0.14
AAC / HE-AAC	aacPlus v2 Content Creation Encoder V7.2.5a aacPlus v2 Content Creation Decoder V7.2.5a
WMA 10	Windows Media Audio 10 Professional BETA evaluation version for EBU
WMA 9	Windows Media Audio 9.1 Professional
Layer 2	MPEG-1 Layer 2 DAB Rm0 SAC Audio Encoder V0.9 MPEG-1 Layer 2 DAB Rm0 SAC Audio Decoder V0.9
MPEG Surround	MPEG Surround Encoder V1.0 MPEG-4 Audio Decoder
MP3 Surround	Surround: MP3s Encoder V04.00.04 MP3s Decoder V04.01.00
DTS	DTS C++ Encoder Version 220.12 DTS Coherent Acoustics Decoder Version 2.30.20 DTS HD Pro Series Encoder v0.97 DTS HD Decoder Library Version 300.38

The table below shows the version number of the various codecs used to generate the test and training sequences in Phase 2.

Table 6: Codec Versions (Phase 2)

Codecs - Phase 2	Version
Dolby Digital	Encoder Model DP569, Firmware version: v.2.0.3.1 Decoder Model DP 562
Dolby Digital plus (Phase 2 new)	Dolby Digital plus Prototype Broadcast Encoder Version 1.6.0 Dolby Digital plus Decoder-Converter Simulation Version 1.1.7
Dolby Digital plus (Phase 1 alt)	Dolby Digital plus Prototype Encoder Version 1.4.0.2 Dolby Digital plus Decoder-Converter Simulation Version 1.0.14
MPEG Surround Layer 2	FhG MPEG-1 Layer 2 DAB Rm0 SAC Audio Encoder V0.9 (build Oct.06) FhG MPEG-1 Layer 2 DAB Rm0 SAC Audio Decoder V0.9 (build Oct.06)
AAC/HE-AAC	Coding Technologies aacPlus v2 Evaluation Encoder V8.0.1 Coding Technologies aacPlus v2 Evaluation Decoder V8.0.1
DTS	DTS HD Pro Series Encoder v0.97 DTS HD Decoder Library Version 300.38

6.2 Verification of bit-rate

Following encoding the resulting files with the encoded bitstreams were used to verify the bit rate for each codec. Verification was carried out for each software codec, each test file and each bit rate. It showed how close the actual bit rate was to the selected (nominal) bit rate.

Bit rates was calculated using the following formula:

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

Verification of bit-rate in Phase 2 was performed in a slightly different way than in Phase 1. Instead of checking the bit rate of each individual test and training item separately, the bit-rates of all test and training sequences were verified as a batch, which made the whole verification process mush faster and more accurate.

It should be pointed out that the Windows-Media encoder used a bit-rate, which often exceeded the selected bit-rate by a bit more 20 kbit/s, (depending on the selected test sequence). This behaviour has to be taken into account, when comparing the results of the different codecs. The following two tables show the actual bit-rate for all codecs in the test and all test and training sequences selected for Phase 1.

Table 7: Actual bit rates for Phase 1 codecs

Item	AAC 256 kbit/s	AAC 320 kbit/s	DD+ 200 kbit/s	DD+ 224 kbit/s	DD+ 256 kbit/s	DD+ 448 kbit/s	DTS 1509 k bit/s	HE-AAC MPEG Surround 64 kbit/s	HE-AAC MPEG Surround 96 kbit/s
Applause	258.1	322.7	200.5	224.7	256.4	449.0	1520.4	66.3	98.4
Bach_organ1	258.1	322.2	200.4	224.5	256.3	448.5	1521.8	66.8	98.6
Bach_organ2	257.8	322.0	200.6	224.7	256.6	449.0	1516.2	66.3	98.6
Exodus	257.9	321.9	200.6	224.8	256.6	449.2	1520.2	66.5	90.3
Hadouk	257.9	322.1	200.5	224.7	256.6	449.0	1516.1	66.4	98.2
Harpsichord	257.8	321.8	200.4	224.3	256.3	448.6	1514.2	66.2	97.9
Hoelsky	258.3	322.1	201.2	225.3	256.9	449.9	1522.7	67.3	99.6
HVB_Gregorian-chant	257.9	321.8	200.3	224.3	256.2	448.4	1513.1	65.9	97.7
Moonriver_Mancini	258.1	322.0	200.8	224.5	256.5	449.1	1519.9	66.3	98.2
R_Plant_Rock	258.0	322.0	200.5	224.4	256.4	448.8	1515.6	66.4	98.4
Sax_Piano	257.9	322.0	200.9	224.8	256.7	449.3	1520.6	66.1	98.4
Sedambonjou_Salsa	258.2	322.1	200.9	224.7	256.6	449.4	1523.2	67.0	98.9
SVT_Female-vocal	257.7	322.0	200.4	224.5	256.4	448.8	1514.6	65.9	98.0
Tschaikowsky_04	258.3	321.9	200.9	224.8	256.8	449.6	1519.0	66.6	98.7

Table 8: Actual bit rates for Phase 2 codecs

Item	L2 MPEG Surround 224 kbit/s	MP3 Surround 192 kbit/s	WMA10 192 kbit/s	WMA10 256 kbit/s	WMA9 128 kbit/s	WMA9 192 kbit/s	WMA9 256 kbit/s	WMA9 448 kbit/s
Applause	223.9	193.0	204.7	272.7	133.0	198.9	264.3	452.8
Bach_organ1	223.6	193.1	194.0	263.1	137.7	205.8	273.5	468.5
Bach_organ2	223.8	192.8	207.9	276.8	135.2	198.4	263.9	459.7
Exodus	224.0	193.0	204.3	272.0	135.0	201.9	263.9	459.7
Hadouk	223.8	192.9	209.5	279.0	135.9	206.8	270.0	471.0
Harpsichord	223.7	192.6	202.5	269.7	134.2	200.4	263.4	457.3
Hoelsky	223.2	193.7	210.9	287.8	138.1	206.1	274.0	481.5
HVB_Gregorian-chant	224.1	192.5	199.5	257.7	132.6	198.0	263.4	451.9
Moonriver_Mancini	224.1	192.6	207.4	276.0	132.9	198.4	263.9	452.2
R_Plant_Rock	223.6	192.8	205.3	273.4	133.7	199.7	265.3	454.8
Sax_Piano	224.0	193.0	206.2	274.9	131.8	197.1	2620	448.9
Sedambonjou_Salsa	223.7	193.2	212.5	277.4	134.9	200.9	267.3	467.1
SVT_Female-vocal	223.7	192.7	198.2	264.1	130.1	194.3	258.6	443.2
Tschaikowsky_04	224.3	193.3	196.8	267.5	135.3	201.4	268.0	449.6

The following table shows the difference between the selected and actually measured bit-rates for each codec under test. The maximum deviation from selected bit-rate was found for the MPEG AAC and in particular for MPEG HE AAC at 128 kbit/s with a deviation of not more than plus 0,625% from the selected bit-rate.

Table 9: Comparison of nominal and actual bit rates

Multichannel Audio Codec	Target bitrate	Measured BR
Dolby Digital (Dolby AC-3) 448 kbit/s (Hardware real time codec, which produced an data-stream instead of a file)	448	No files available
Dolby Digital plus (new version, used only for Phase 2)	200	200.0
Dolby Digital plus (new version, used only for Phase 2)	256	256.0
Dolby Digital plus (old version, which had been used in Phase 1)	200	200.0
Dolby Digital plus (old version, which had been used in Phase 1)	256	256.0
MPEG Surround Layer 2	256	256.4
MPEG AAC 320 kbit/s (old version, which had been used in Phase 1)	320	321.6
MPEG HE AAC 128 kbit/s	128	128.8
MPEG HE AAC 160 kbit/s	160	160.8
MPEG HE AAC 192 kbit/s	192	192.9
Transcoding: Dolby Digital plus 256 kbit/s → DD 640 kbit/s (new version of DD plus, used only for Phase 2)	640	640.0
Transcoding: MPEG AAC 320 kbit/s → DTS 1500 kbit/s	1509	1509.0
Transcoding: MPEG HE AAC 192 kbit/s → DTS 1500 kbit/s	1509	1509.0

7. Experimental design

7.1 Test methodology

The test method adopted by B/MAE Group uses experience gained by EBU Project group B/AIM with the so-called MUSHRA method ("MUlti Stimulus test with Hidden Reference and Anchors", ITU-R BS.1534) that supplies a successful approach for assessing and grading different impairments. The scale for the grading was based on a video signal evaluation method (ITU-R BT.500). A quality scale is used where the intervals are labelled "bad", "poor", "fair", "good" and "excellent" as opposed to ITU-R BS.1116 which uses an impairment scale. The value on the lower end of the scale is zero; the value on the upper end is 100. No decimals are given. The method uses the unprocessed original programme material with full bandwidth as the reference signal. The set of processed signals consists of all the signals under test and three additional signals (anchors).

The B/MAE Group has made a decision to use the MUSHRA methodology for all tests as a main evaluation technology. MUSHRA covers the whole quality range and is easier to run than BS 1116.

Initially, BS 1116 was to be used for very high quality sequences graded above a given threshold (e.g. 80 points on MUSHRA scale), but during the tests it was considered that

this methodology was not necessary, as MUSHRA provided enough discrimination.

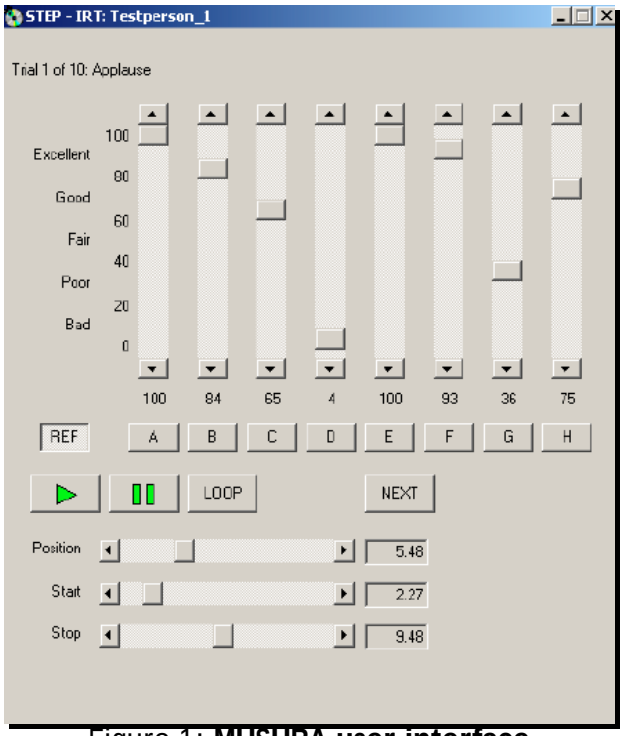


Figure 1: **MUSHRA user interface**

The figure above shows the graphical user interface of MUSHRA STEP software package.

Evaluation software

All but one laboratory used the same Skylab Multichannel Audio Test system STEP software. The STEP interface used is shown on figure below.

The exception was France Telecom – Orange who used the CRC SEAQ test package.

Impairment (artefact) categories for MCA tests

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

Table 10: Explanation of the artefact categories

No.	Artefact category	Explanation
1	Signal correlated noise	coloured noise associated with the signal
2	Loss of High Frequency	lack of high frequencies, dull sounding
3	Excess of High Frequency	excess of high frequencies or associated effects,e.g. sibilance or brightness
4	Periodic Modulation Effects	periodic variations such as warbling, pumping, or twitter
5	Temporal Distortion	pre- and post-echoes, smearing, effects associated with transients
6	Distortion	harmonic or inharmonic distortion
7	Loss of Low Frequency	lack of low frequencies
8	Image Quality	all aspects including narrowing, spreading, movement and stability
9	High frequency distortion	distortion in the high frequencies, including phase distortions

7.2 Anchors

The choice of appropriate anchors is fundamental both for subject rejection and for statistical issues (such as test labs comparison). The MUSHRA methodology was basically developed in order to test stereophonic audio sequences. That is why it needs some adaptation to be used for multichannel audio testing. From those considerations, the choice of the anchors was the following:

- A hidden reference (unprocessed signal)
- A low anchor signal: a filtered version (3.5 kHz low pass) of the unprocessed signal.
- Spatial anchor signal: generated by introducing deliberate crosstalk between the channels, resulting in the distortion of the spatial image.

The two first listed anchors are mandatory by the MUSHRA methodology. The spatial anchor has been specifically defined for our multichannel tests.

7.3 Evaluation Process

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

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

In PART B of the training phase the subject was able to listen to all 4 multichannel audio training items at the different possible degradations in order to illustrate the whole range of possible qualities. Similar to the test items, these training items were more or less critical depending on the bit rate and other "conditions" used. In this part of the training phase the subject was asked to use the available scoring equipment and to evaluate the quality of the items by inputting the appropriate scores on the Quality Scale. For each multichannel audio item, the explicit reference and the “codecs under test”, which include the hidden reference and two other anchors items, were presented.

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

A copy of the test instructions used by English speaking assessors is given in **Appendix 2**.

The purpose of the grading phase is to score the items across the given quality scale. The subject scores should reflect subjective judgement of the quality level for each presented multichannel audio items. Each trial will contain up to 9 multichannel audio signals to be graded. Each item is approximately 20 seconds long.

The subject should listen to the reference and all the test conditions by clicking on the respective buttons. It was allowed to listen to the signals in any order, any number of times. Repeated playback is available thanks to a “Loop” button. If the subject wants to focus on a certain part of the multichannel audio item, he is allowed to select thisby changing the start and end markers. The subject is allowed to adjust these markers only after listening to the whole multichannel audio item. A slider for each signal was used to indicate the subject’s opinion of the current signal

quality. Once the subject is satisfied with his grading, he clicks on the “Trials” or “Next” button at the bottom of the screen in order to get the next trial.

The following Quality Scale was used:

- Excellent
- Good
- Fair
- Poor
- Bad

The scale is continuous from "Excellent" (100) to "Bad" (0).

When evaluating the items, the subject does not necessarily give the grade "Bad" to the item with the lowest quality in the test. However one or more items must be given the maximum grade of100 because the unprocessed reference signal is included as one of the multi-channel audio items to be graded. During the training phase, the subject should be able to learn how to interpret the audible impairments in terms of the grading scale and he is encouraged to discuss personal interpretation with the other subjects at any time during the training phase. No grades given during the training phase will be taken into account in the true tests.

7.4 Listening conditions

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

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

7.5 Test Sessions

The listening tests had to provide sufficient statistical coverage of each of the codecs under test, aiming for at least 15 listeners per codec. The number of codecs each listener should listen to was chosen to be 5, which along with the two anchors and a hidden reference gave 8 stimuli along with the known reference. This was considered a sensible number of stimuli for each listener: not too many to cause fatigue or confusion, but enough to get sufficient coverage.

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

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

The codecs were split into three groups consisting of low, medium and high bit rate codecs. Each listener would get one low, one medium and one high bit rate codec plus two more of any of the three groups. This ensured a listener would be exposed to one codec they would find noticeably degraded in the low bit rate group that they would be expected to score below 100. They would also experience a range of qualities with the medium and high rate codecs. By having 3 codecs in the same group it would also benefit the scoring by the same listener assessing the more subtle difference between more closely matched codecs. Of course, grouping vastly differing codec algorithms purely on bit rate is rather crude, but it is at least unbiased towards any particular codec design.

Each session file was generated pseudo-randomly, so that each listener received a different file. It was constrained according to the low, medium and high bit rate groupings just described. The

codecs covered by all the session files were counted to ensure each codec was covered by at least 15 listeners.

The session files were generated for use with the ARL STEP software, which the majority of the test labs were using. These files were ASCII text files containing a list of the required file combinations for each test item. The files were generated by a simple C program that contained a suitable pseudo-random number generating algorithm (based on a modulo-prime counter to give an even distribution) that labeled each audio file used accordingly.

The ARL STEP software automatically randomizes the order of the codecs and test items, so this did not have to be dealt with in the session files.

Phase 1

Phase 1 contained 19 codecs, with 6 participating labs each supplying at least 12

listeners. The codecs under test were split into three groups according to bitrate (5 low, 8 medium and 6 high).

Table 11: Codec groupings in Phase 1

Group 1	Group 2	Group 3
HE AAC MPS 64	DD+ 200	AAC 320
HE AAC MPS 96	DD+ 224	DD 384
MP3 S 192	L2 MPS 224	DD+ 448
WMA10 192	AAC 256	DTS 448
WMA9 192	DD+ 256	WMA9 448
	Prologic2 L2 256	DTS 1,5
	WMA10 256	
	WMA9 256	

Phase 2

Phase 2 contained 13 codecs, with 6 participating labs each supplying at least 6 listeners. The 13 codecs were split into four groups according to bit rate and concatenation.

Table 12: Codec groupings in Phase 2

Group 1	Group 2	Group 3	Group 4
HE AAC 128	DD+ 200	AAC 320	HE AAC 192&DTS 1500
HE AAC 160	DD+ 200 ev	DD 448	DD+ 256&DD+ 640
HE AAC 192	DD+ 256		AAC 320&DTS 1500
	L2 MPS 256		
	DD+ 256 ev		

8. Statistical Analysis and Postscreening

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

The range defined by the confidence intervals is therefore appropriate to give an estimate of the significance of mean differences when comparing different results. In other words from overlapping of confidence intervals it can be concluded that the considered means don't differ significantly. On the other hand from non-overlapping it can be concluded that the means differ significantly.

In order to guarantee reliable test results in subjective tests like these post-screening of subjects is necessary. Three rejecting criteria were acquired. The listeners' results were checked with these

rejection criteria, and those that failed any of these were removed from the results. The three criteria were:

1) Spearman Rank Correlation below 0.8. The individual's scores were ranked and compared to the overall rank of all the listeners. If the correlation of their ranking with the overall listener population was worse than 0.8 they were rejected.

2) If a listener's average scoring for the 3.5kHz anchor was 20 points greater than the overall average they were rejected.

3) If a listener's average scoring for the hidden reference was 20 points less than the overall average they were rejected.

Phase 1

Table 13: Rejected and accepted listeners for each lab (Phase 1)

Lab	Accepted	Rejected
A	17	6
B	10	3
C	20	0
D	9	3
E	15	2
F	16	2

Phase 2

Table 14: Rejected and accepted listeners for each lab (Phase 2)

Lab	Accepted	Rejected
A	2	7
B	16	2
C	13	2
D	9	5
E	12	0
F	5	0
G	14	0

9. Presentation of Main Results

This section presents the aggregate results for Phase 1 and Phase 2, as follows:

- All codecs averaged over all test items
- Average scores for all items over all codecs
- Average scores for "Applause" item for codecs common to Phases 1 and 2

The detailed results for each codec tested over all test items are given in **Appendix 4.**

9.1 Phase 1

All Codecs averaged over all test items plus average of worse case item.

The black line represents the mean over all items. The grey box is the 95% confidence interval for that mean. The lower point of the vertical red bar is the mean of the worse case item.

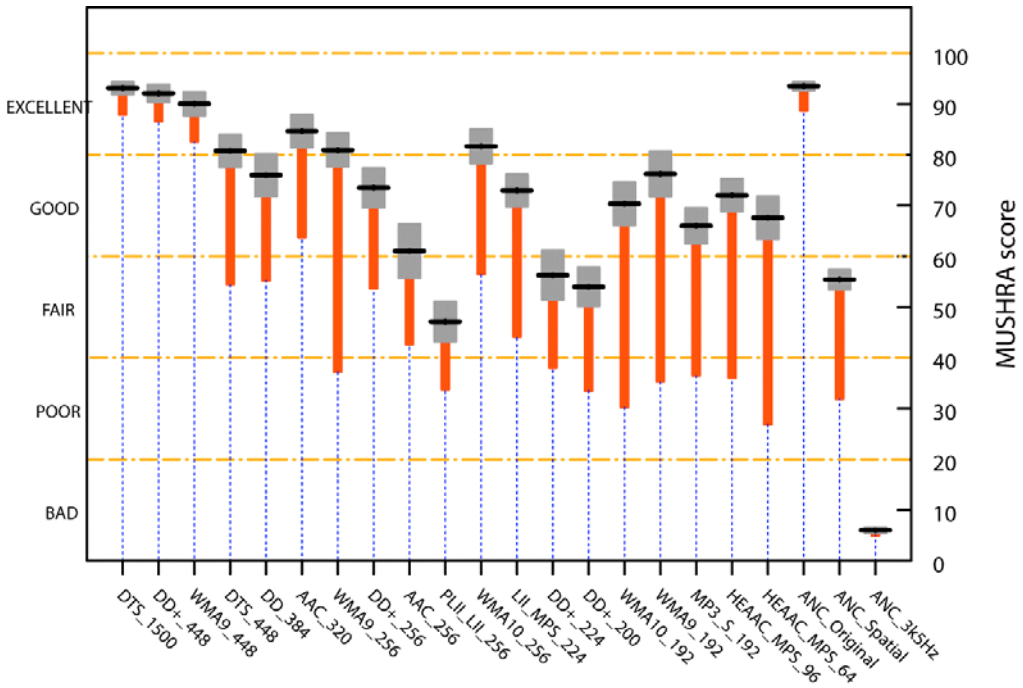


Figure 2: MUSHRA scores of all codecs averaged over all test sequences in Phase 1.

black horizontal lines: mean values
gray areas: 95% confidence intervals
bottom of red bars: mean values of the worst case item

Average scores for items over all codecs

This compares how critical each test item used was by taking the mean for each item over all the codecs used.

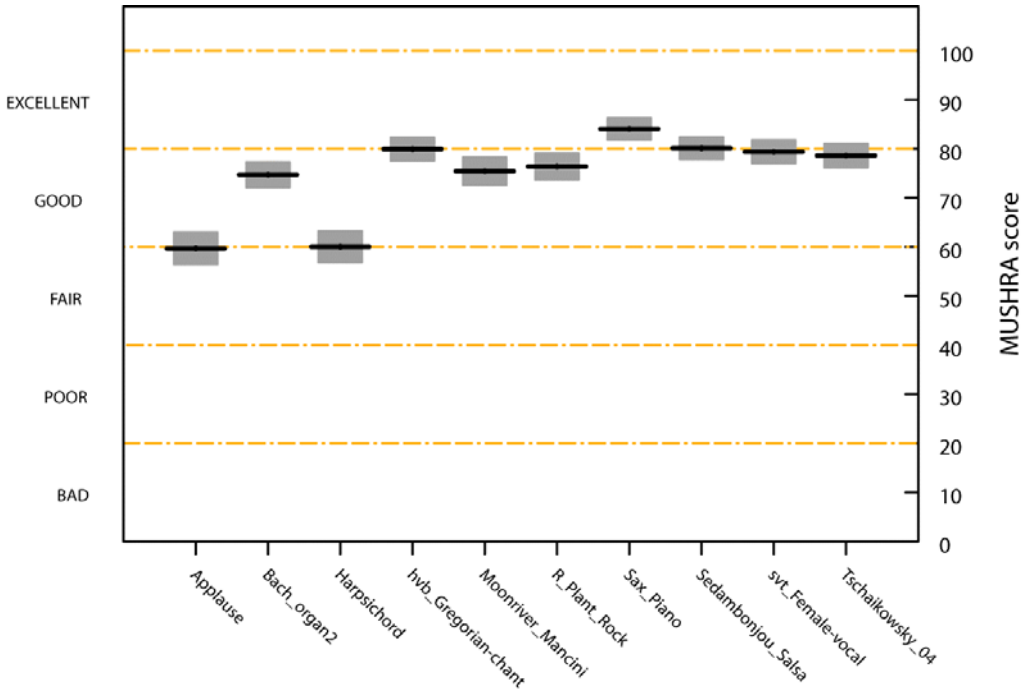


Figure 3: MUSHRA scores of all items averaged over all codecs in Phase 1

black horizontal lines: mean values
gray areas: 95% confidence intervals

9.2 Phase 2

All Codecs averaged over all test items plus average of worse case item.

The black line represents the mean over all items. The grey box is the 95% confidence interval for that mean. The lower point of the vertical red bar is the mean of the worse case item.

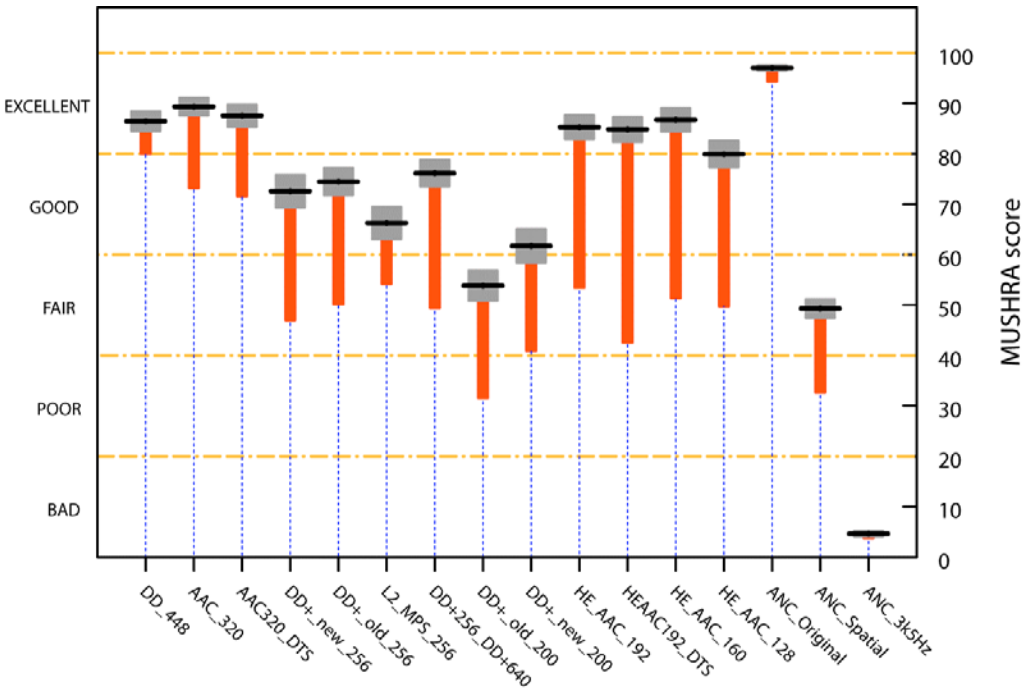


Figure 4: MUSHRA scores of all codecs averaged over all test sequences in Phase 2

black horizontal lines: mean values

gray areas: 95% confidence intervals
bottom of red bars: mean values of the worst case item

Average scores for items over all codecs

This compares the how critical each test item used was by taking the averages for each item over all the codecs used.

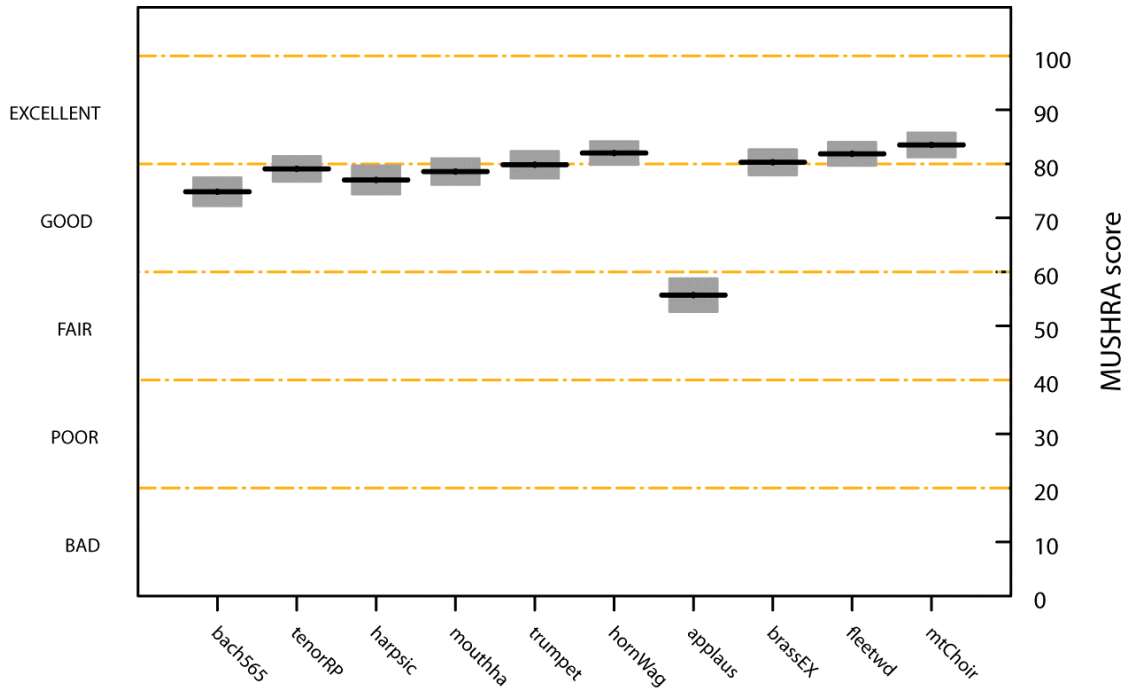


Figure 5: MUSHRA scores of all items averaged over all codecs in Phase 2

black horizontal lines: mean values
gray areas: 95% confidence intervals

9.3 Phase 1 and 2 Comparison

Average scores for Applause item for codecs common to phases 1 and 2

This compares the results from Phase 1 and Phase 2. Only the “applause” item was common to both tests. The common codecs were AAC 320, DD+ 200 (old version), DD+ 256 (old version) and the three anchors. The results in Figure 5 show that the confidence intervals in each case overlap regarding the codecs from Phase 1 and the same ones in Phase 2. That means that there are no significant differences between the results form Phase 1 and Phase 2 and consequentially a high degree of reliability and comparability could be achieved.

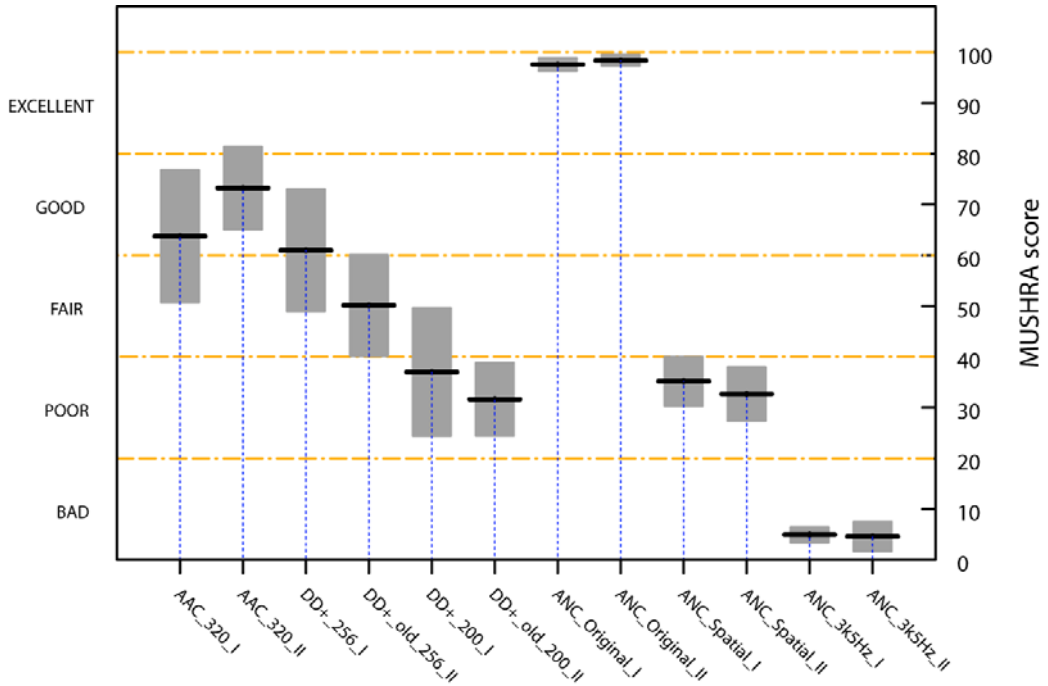


Figure 6: MUSHRA scores of codecs used in both Phase 1 and Phase 2 for "applause"

10. Summary and Conclusions

Nearly ten years after having performed the first large-scale evaluations of early multichannel audio codecs, the EBU members embarked into a new test campaign, this time with even more significant effort and more codecs involved. During the past ten years a lot of new multi-channel audio codecs have appeared in the market. This is particularly true for the last couple of years. Many of these new codecs are already included in the DVB toolbox and standardised by ETSI. Many of them have been only subjected to some internal and informal tests rather than being thoroughly tested by independent authorities. To this end, the users have been confused as to which codec is most suited for different broadcast applications.

For this reason, the EBU Broadcast Management Committee decided about a year ago to launch comprehensive tests to obtain detailed information about the performance of all major multi-channel audio coding systems that might be of interest for broadcast applications. The plan was to have two rounds of subjective tests, one concentrating on all emission codecs, and the other concentrating on contribution and primary distribution codecs, including multiple encoding and decoding and cascading with some of the most promising emission codecs.

However, due to a slight problem with the HE AAC codec discovered by the codec proponent just before the start of subjective testing, the first round had to be split into two phases. The first phase could not include the faulty HE AAC codec. As the bug was fixed quickly, the HE AAC codec could then be included in the second phase, along with the improved versions of Dolby Digital Plus and the two cascaded codecs (i.e. Dolby Digital Plus - Dolby Digital and MPEG-4 HE AAC - DTS). The cascades are entirely new development and have not yet been tested in the international arena. They are potentially of interest to broadcasters and end users because they are compatible with existing multi-channel audio home receivers.

The EBU is planning to perform a third round in order to test the performance of codecs for postproduction and primary distribution. In addition, we plan to assess the performance of several multiple-cascaded multi-channel audio codecs. It has been shown that in practice a broadcast chain may consist of several (5 or more) different codecs, and the overall chain performance can be significantly degraded. This calls for additional tests, which EBU plans to carry out in the near future.

The effort of EBU member organisations that was going into these tests is remarkable. The subjective testing was conducted at eight different EBU test laboratories: Bayerischer Rundfunk, BBC Research, Orange France Telecom, ORF, IRT, Radio France, RAI CRIT and TVP. During the first test round, over 14000 scores were produced by approximately 150 test subjects. The most significant part of these tests was taken by IRT that performed the preparation for the two phases of these tests, including the encoding of thousands of critical items, the selection of the final test and training items out of more than 100 multichannel audio sequences as well as the final processing of the items to be included in the MUSHRA test system.

The results of each assessor were scrutinized and compared to the average value of all assessors. Those assessors who deviated too much from the average were discarded from the tests. Their scores were not used for further statistical analysis of the results.

The general conclusion of the EBU evaluations is that the quality performance cannot be achieved if the bitrates used are not sufficient. This conclusion applies to both old and new codecs. If the quality performance requirement for broadcasters is that none of the test sequences resulted in a quality lower than “Excellent” (i.e. 80 points on MUSHRA scale), then relatively high bitrates are required. For example, consider Dolby Digital (DD) or DTS which have been in the market for more than 10 years: Dolby Digital requires 448 kbit/s and DTS still requires around 1.5 Mbit/s for "Excellent" quality. The newer codecs, such as Dolby Digital Plus or Windows Media provide "Excellent" quality only if operating at 448 kbit/s or above.

It is interesting to note that that broadcasters who were using Dolby Digital at a bitrate of 448 kbit/s several years ago made the right decision, although it was not made on scientific basis but was based merely on practical "trial-and-cut" experience. Today, Dolby Digital still represents a good compromise between bit rate and quality. Broadcasters who use DD at 448 kbit/s for 5.1 multi-channel audio are able to offer excellent multi-channel audio quality. This conclusion is equally true for standard TV, HDTV and radio broadcasts.

The MPEG AAC codec operating at 320 kbit/s shows an “Excellent” quality level on average, however it performs less well for “applause” (i.e. “Good”). Unfortunately, AAC could not be tested at 448 kbit/s, therefore a direct comparison with Dolby Digital and Windows Media is not possible at that bitrate.

Nevertheless, HE AAC codec gives really remarkable results. For bitrates equal and higher than 160 kbit/s, the average of all ten test items was found to be in the region of "Excellent". This means that the mean value of HE AAC is similar to the mean value of the above mentioned codecs operating at almost 3 times higher bitrate! On the downside, HE AAC gives a rather unbalanced behaviour, as "applause" is always in “Fair” region, independently of the selected bitrate.

Equally remarkable is the quality performance of HE AAC at 128 kbit/s. In spite of extremely low bitrate (for multichannel audio!), it scores systematically between “Good” and “Excellent”, with the exception of "applause" which is again in the “Fair” region only.

It can be concluded that, at the moment, the MPEG HE-AAC seems to be the most favourable choice for a broadcaster requiring a good scalability of bitrate versus quality, down to relatively low bit rates. In addition, the AAC-based codec family offers excellent audio quality at higher bitrates, e.g. at 320 kbit/s (with the exception of "applause"). Our study shows that excellent quality (on average) can be achieved even at half the bitrate, i.e. 160 kbit/s, or even less, for all test items except for the most critical items.

The new coding systems with parametric coding of the surround information of the various audio channels, such as MPEG Surround, show a rather unbalanced behaviour which depends on the type of the test sequence. For the MPEG Surround codecs "applause" is again the most critical sequence resulting in only "Fair", or sometimes even "Poor" audio quality. It can be concluded that the MPEG Surround codecs at the moment do not fulfil the requirements for high quality broadcasting at their target bit rate, or do not offer the expected advantage in terms of bitrate gain compared to already well-established codecs. These codecs might be appropriate for rendering of 5.1 multi-channel audio which is broadcast, for example, via narrow-band DAB channel with acoustic

constraints, such as 5.1 multichannel audio in a car or in other noisy environments, and still maintaining backwards compatibility with two-channel stereo receivers in order to avoid simulcasting of stereo and 5.1 multichannel audio signal. For these cases, the new MPEG Surround standard could be a good candidate.

In general, the results of this EBU study show that multichannel audio codecs made a significant progress since our last evaluation campaign about ten years ago. Nevertheless, excellent multi-channel audio quality, which is generally required by broadcasters, cannot always be achieved; either because of low bitrates used (due to narrow band transmission channels used) or because of some particular unfriendly content (such as "applause" in the case HE AAC). In order to achieve the quality in the "excellent" region for any content and regardless the codec used, bitrate of 448 kbit/s is a safe choice. If, however, bitrates below 448 kbit/s are used, compromises concerning the quality may have to be accepted.

It is very important that the service providers know exactly the applications of their 5.1 surround sound services and understand the limits of their services. Broadcasters endeavour to preserve high quality of their broadcasts for most of the time. This is unfortunately not always possible. There may be some critical audio sequences for a portion of time that are reproduced with poor quality, even though the average quality for most of the time is good. In many cases it may be wise to preserve intrinsic audio quality by delivering a high quality two-channel stereo signal (which requires less bandwidth) instead a compromised 5.1 surround signal. If there is a serious constraint in terms of bandwidth, so that a broadcaster is advised to use lower bitrates, it is often a better strategy to deliver a good stereo downmix signal to the end user than poor or even bad multichannel audio signal.

As broadcast mechanisms are one of the main sources of discrete surround sound in the consumer environment, outmost care should be taken by the broadcaster to deliver near-studio-quality to the intended audience. The EBU audio experts are of the opinion that the bitrate budget for broadcast multichannel audio should not be reduced in any way, unless strictly necessary, as this may severely degrade the end-user experience and may not meet high expectations associated with new-generation digital broadcasting, including HDTV.

11. Acknowledgements

The B/MAE Project Group would like to gratefully acknowledge the work and assistance of all the EBU laboratories participating in the tests: BBC, IRT, TVP, FT Orange, Radio France, RAI CRIT, BR and ORF. Particular thanks should go to all listeners (more than hundred!) who voluntarily participated in our tests. We should also thank the system proponents Dolby, DTS, Coding Technologies and Microsoft for their support. And finally, the work of the two selection panels who made the selection of critical sequences should be duly acknowledged. The Group appreciates the work of Gerhard Spikofski for his effort in providing the complex statistical analyses and David Marston who was responsible for setting up the session files.

12. References

[1] EBU, “BPN 019: Report on the EBU Subjective Listening Tests of Multichannel Audio Codecs”, March 1998 *Available to EBU members only.*

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

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

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

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

Appendix 1: Selection Panel Report for Phase 1 (June 2006)

Description of process

The aim of the process was to find 10 items of multi-channel audio material for use in subjective tests. Based on previous experience, rather than explicit instruction, the panel also selected 4 additional items to be used for training for the tests.

There were a number of important considerations, documented in ITU-R Recommendation BS.1116. The material chosen should not be wearisome, nor too involving, and should be normal programme material. The chosen test items should be those that reveal something about the performance of the systems under test, including differences between systems. There should be a range of different types of material and care must be taken to ensure that there is no bias towards or away from any particular system.

Submission and coding of candidate material

All parties involved in the B/MAE group were invited to submit candidate items. The submissions were made to IRT. In all, about 60 were received. Most items were of an ideal length for testing (between 10s and 30s). A few significantly longer items were included too. Members of staff of IRT performed coding and decoding. A total of 13 different codecs were used, some software, one real time hardware.

Presentation of coded items

The identity of the codecs was not revealed during the selection panel process. Items were re-played from a PC operated by a member of IRT staff. Codecs were identified only by a letter ("A", "B", "C", and so on). There were some codecs that were included at 2 bit rates. It was felt useful for the selection panel to know this, and so some codecs were identified as, for example, "C1" and "C2" to indicate codec "C" at two different bit rates. The mapping of codec to identifier was known only to the operator of the replay system.

Auditioning and decision-making

The panel listened to all combinations of proposed test item and codec - a total of approximately 650 combinations. During the listening process the panel noted observations for each combination and discussed them. The PC replay system allowed the panel to request that any combination be repeated, including the original at any time. Where items were longer than is suitable for a test, a representative sub-section was chosen for detailed listening following a brief audition of the complete item.

For convenience, a set of impairment types was used as the basis for discussion. This is shown in Table below. This table is derived, with changes, from ISO/IEC JTC1/SC29/WG11 No 685, March 1994. It is the same as was used for the EBU B/CASE multi-channel audio codec tests in 1998.

Impairment categories are given in Table 20 of this report.

After the initial listening the observations about all items and codecs were discussed at length. The first step was to dismiss those that did not reveal any significant artefacts. A small number that failed the "wearisome" test were also removed. Of those that remained, some would clearly be very useful and some had potential. Fortunately, there was no shortage of "good" items. A short list of 20 was made with no great difficulty - taking carefully into account the range of types of impairments each item evoked, and the range of codecs that suffered from those impairments.

The 20 short-listed items were auditioned again in the listening room to verify the range of programme material and the types of impairment observed. Where there were several items of similar nature some were removed from the list, and some were separated into a "test" and a "training"

item. Two lists, one of ten items for actual tests and on of 4 items for training of subjects, were produced.

Table 15: Items selected for test sessions

No.	Name	Description
1	Applause	Applause with distinct clapping sounds
2	sax_piano	saxophone and piano
3	R_Plant	rock music (“Whole lotta love”)
4	Tschaikowsky_04	orchestral piece; open sound
5	moonriver_mancini	mouth organ and string orchestra
6	sedambonjou	atmospheric performance of Latin-American music
7	Harpsichord	solo harpsichord; isolated notes
8	svt	live performance of female vocalist and band; crowd sounds
9	HVB	small choir; large church; Gregorian chant
10	Bach_13_5	church organ; lots of stops out

Table 16: Items selected for training sessions

No.	Name	Description
1	hadouk	Eastern woodwind, strings, percussion
2	Exodus	orchestral; lots of brass instruments
3	Hoelsky	“chattering” choral voices
4	Bach_13_3	church organ; few stops out

Table 17: Categories of impairment for all codecs and test items

Item	A1	B1	C1	C2	D1	D2	E1	F1	F2	G1	H1	I1	I2
Applause	5	2	2,5	2,5	1,9	1,9	5	2,5	2,5	5	5	5,9	2,5
Hadouk	4,6, 8	5,8, 9	5	5	2,6	2,6	3,5, 8	1,8	6	3,9	3	5	4,9
R_Plant	5,9	7,9	2,5, 8	s,5, 8	2,7, 8	2,5, 8	5	2,5, 9	2,5, 9	5	8	5,9	5
Sax_piano	2,8	7,8	5,8	2	2	2	2,8	6,8	4,6	2,6, 8	6	6,9	
Moonriver_Mancini	2,6	8	6	4,5, 6	4,5, 6	4,8	4,6	6	6	1,2, 6	1	6,8	6
Exodus	9	8	2,8	8	2,6		9	6	6	8,9	5	6,8	2,6
Hoelsky	2,8	2,8	1		6	6	8	5	6	8	6,8	6,8	6,8
Sedambonjou	2,5	6,8	2,3, 6	2,6	2,5	2	7	6,9	6,9	2,6	1	2,5	6
Tschaikowsky_04	6	6	8	2,6	8	5,7	6	6	6	6	3,6	6	6
HVB	3,8	4,5, 8	2,4, 6	4	6,8	4	6,8	4	4	4,6	1,6	6	6
svt	5	5	1,5	1	9	6	3	6	3	2,4, 8	6	4,9	4,9
Bach_13_3	2	6	1	1,8	2,8	1,6	1,6	6	6	6	6	4,6	4,6
Bach_13_5	2,8	6,8	8	8	2,6, 8	8	6	6,8	6,8	6,8	2	6	6
Harpsichord	1,2, 5	1,6	2,4,5, 6	4,5	2,4	2,4	1,5, 6	1,4, 6	4	1,4	2,4, 5,8	1,2, 5,6	1,2, 5,6

Description of listening facilities

The selection panel used a listening room (Room AE14) at IRT, Munich, for listening. The room is purpose-designed for listening. The loudspeaker arrangement meets ITU-R BS.775 with a stereo basis L – R and loudspeaker distance of 3m. The Type of loudspeakers are 5 MEG RL922 (Musikelectronic Geithain) and 2 subwoofers (1094A and 1092A Genelec), where the 1094A radiates the summed low frequency ratio below 80 Hz of the three front channels and the 1092A the corresponding low frequency ratio of the two surround channels and additionally the LFE signal. The listening situation meets the recommendations EBU doc. Tech 3276 including EBU doc. Tech 3276 (Supplement 1) as well as ITU-R Recommendation BS.1116.

Replay from the PC used a digital soundcard, through a digital routing matrix (DAIS) into a Yamaha O2R mixing console. The bus outputs of the console were fed to Yamaha DACs then to the above mentioned loudspeakers.

Seven seats were arranged in two rows. The middle seat of the three in the front two was at the sweet spot. The rear row of four was placed centrally, immediately behind the front row. It was accepted that the listeners at the ends of the back row would hear a sound balance that was skewed toward one or other of the surround loudspeakers. The possibility of having the panel members change place whilst listening to a particular item was dismissed as impractical given the size of the panel and the number of items to be auditioned. Each panellist used the same seating position throughout.

In the opinion of the panel, the equipment was more than adequate for the task.

Problems encountered during the selection process

One of the items, "hadouk", triggered extreme behaviour in codec D1. Very severe impairments were produced in parts of the item. The codec was assumed not to be completely broken because D2, known to be related, reacted in a similar, but much more controlled way. The decision was made to retain this item as a training item. The proponent of the codec will be made aware of the behaviour on completion of the tests.

Two of the coders, D1 and D2, failed to encode the original "moonriver_mancini" file, producing an uninformative error message immediately. The team performing the encoding observed that this file was the only 16 bit file in the set. Converting it to 24 bit made it acceptable to the D1 and D2 coders.

Reproduced Sound Level

The selection panel auditioned each of the 14 selected items and agreed an appropriate sound level for the replay of each, taking into account comfort, and programme type. There was some disagreement around the "R_Plant" item when it was found that the volume control only went up to 10.

The mixing console fader settings used for each item were noted and will be used in combination with the pre-encoding alignment adjustments to adjust the level of decoded items. In this way, all files in the subjective tests will be replayed at the appropriate level following a simple initial calibration.

The following table gives the necessary alignment adjustments which will be done before the coding process (QPPMmax = -9dbFS, level meter with 10ms attack time), and the listening level adjustments which are naturally done after the decoding process to achieve the appropriate listening level during the subjective test procedure.

In order to reproduce the correct listening level a defined measuring signal (band limited pink noise with a defined signal level) will be sent to each test lab together with the en-/decoded test items.

Gain adjustments required for coding and listening level

Table 18: Gain adjustments

Item	fader setting during selection	alignment gain change required before coding	fader setting expected after alignment	gain change applied after decoding	resulting listening level (SPL, dBA)
bach_13_5	-15.8	2	-17.8	-11	71
applause	-12.9	-2.8	-10.1	-4	78
tschaikowsky_04	-14	-7.5	-6.5	0	82
sax_piano	-16	-5	-11	-5	78
harpsichard	-11.6	13	-24.6	-18	64
sedam bonjov	-19.6	-5.7	-13.9	-7	75
moonriver-mancini	-11.6	6.7	-18.3	-12	70
r-plant	-17	-2.8	-14.2	-8	74
svt	-12.9	0.7	-13.6	-7	75
hbv	-18.9	1.5	-20.4	-14	68
bach_13_3	-15.8	10	-25.8	-19	63
exodus	-18.9	-4.3	-14.6	-8	74
hadouk	-20	-4.5	-15.5	-9	73
hoelsky	-20	-6.1	-13.9	-7	75

The correct listening levels for all processed items will be achieved by setting the SPL of reproduction of band-limited pink noise from the test DVD "Multichannel Universe" (GLS-9500-3) to 67dBA for each loudspeaker, which should give 74dBA for 5 loudspeakers.

Appendix 2: Instructions for Assessors

Subjective testing of multichannel audio coding systems

Thank you for participating in a subjective evaluation of multichannel audio coding systems.

Background

Audio coding systems are used to reduce the amount of data required to represent an audio signal. The reasons can be to reduce storage requirements, transfer time, or bandwidth requirement. A couple of new multichannel audio coding systems appeared on the market recently, which are of interest for video as well as audio broadcast applications.

The aim of these tests it to measure the effect on audio quality of the different, commonly used, multichannel audio coding and decoding operations, at a variety of data rates typical of broadcast use.

You are asked to assess basic audio quality, taking into account **all aspects** of the audio signals.

For your convenience a set of impairment types in the table 1 below can be used as the basis for your quality evaluation. This table is derived, with changes, from the MPEG document No. 685, March 1994. It is the same as was used for the EBU B/CASE multi-channel audio codec tests in 1998.

Table 19: Impairment categories

No.	Artefact category	Explanation
1	Signal correlated noise	coloured noise associated with the signal
2	Loss of High Frequency	lack of high frequencies, dull sounding
3	Excess of High Frequency	excess of high frequencies or associated effects, e.g. sibilance or brightness
4	Periodic Modulation Effects	periodic variations such as warbling, pumping, or twitter
5	Temporal Distortion	pre- and post-echoes, smearing, effects associated with transients
6	Distortion	harmonic or inharmonic distortion
7	Loss of Low Frequency	lack of low frequencies, i.e. missing of “basse fondamentale”
8	Image Quality	all aspects including narrowing, spreading, movement, stability and audio objects remaining in their original position
9	High frequency distortion	distortion in the high frequencies, including phasey distortions

Training phase

The first step in the listening tests is the familiarisation with the listening tests process. This phase is called a training phase and it precedes the true evaluation phase.

The purpose of the training phase is to allow you, as an evaluator, to achieve two objectives as follows:

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

In PART B of the training phase you will be able to listen to all 4 multichannel audio training items that have been selected for the training in order to illustrate the whole range of possible qualities. Similar to the multichannel audio test items, those training items will be more or less critical depending on the bit rate and other "conditions" used.

In this part of the training phase you will be asked to use the available scoring equipment and evaluate the quality of the items by inputting the appropriate scores on the Quality Scale. For each multichannel audio item, you will find the explicit reference and the “codecs under test”, which include the hidden reference and two anchor items.

Grading

The purpose of the grading is to score the items across the given quality scale. Your scores should reflect your subjective judgement of the quality level for each of the multichannel audio items presented to you. Each trial will contain up to 9 multichannel audio signals to be graded. Each of the items is approximately 15 seconds long. You should listen to the reference and all the test conditions by clicking on the respective buttons. You may listen to the signals in any order, any number of times.

In order to repeat the playback you are allowed to use the “Loop” button. If you want to focus on a certain part of the multichannel audio item, you are allowed to select a special excerpt by changing the start and end markers. You are allowed to adjust these markers **only after listening to the whole multichannel audio item**.

Use the slider for each signal to indicate your opinion of its quality. When you are satisfied with your grading of all signals you should click on the “Trials” or “Next” button at the bottom of the screen.

The following Quality Scale is used:

Excellent
Good
Fair
Poor
Bad

The grading scale is continuous from "Excellent" to "Bad".

In evaluating the items, please note that you should not necessarily give grade "Bad" to the item with the lowest quality in the test. However one or more items must be given the grade Excellent (100) because the unprocessed reference signal is included as one of the multi-channel audio items to be graded.

During the training phase you should be able to learn how you, as an individual, interpret the audible impairments in terms of the grading scale. You are encouraged to discuss your personal interpretation with the other subjects at any time during the training phase. No grades given during the training phase will be taken into account in the true tests.

Important remarks for the test supervisor

- After the adjustment of the appropriate listening level by an individual subject, the adjusted listening volume has to be noted for each subject in order to guarantee always the same appropriate listening volume for this subject.
- The supervisor shall not influence the subject concerning the weighting of the different quality-parameters

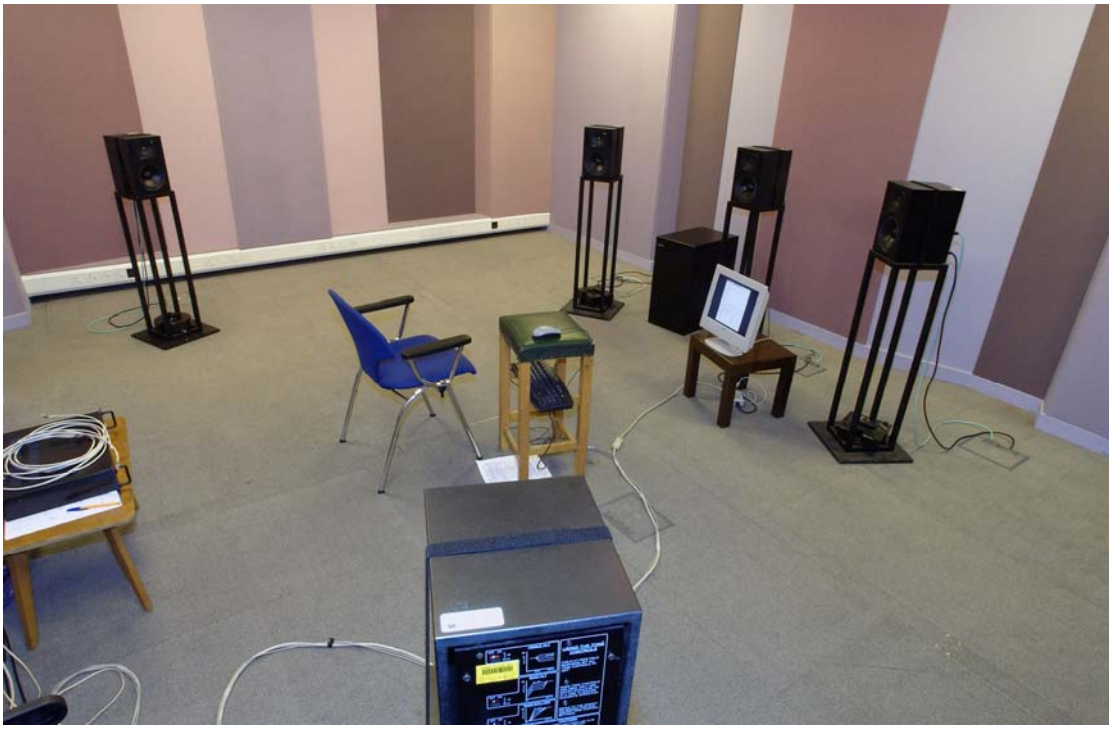
Page intentionally left blank. This document is paginated for recto-verso printing

Appendix 3: The EBU Members' Listening Rooms and Equipment Set-up

This chapter describes the listening rooms and the equipment set-up of the EBU laboratories involved in the tests. The following laboratories contributed the relevant data:

- BBC
- France Telecom - Orange
- IRT
- ORF
- Radio France
- RAI CRIT

British Broadcasting Corporation (BBC), Kingswood Warren, UK



Size:

Basic shape .....rectangular

Floor area .....5.3m x 4.7m

Height 2.7m

Acoustics:

Basic room treatment: Controlled reflection (R. Walker room)

Mixed absorption (rock wool) and hard reflective surfaces

Loudspeakers:

Type (manufacturer): 5x Genelec 1030A (set to “flat”)

1x Genelec 1092A subwoofer (set to roll off –2dB)

Technical equipment:

Audio workstation:

Hardware: Windows 2000 PC

Audio card: Lynx TWO B

Software: ARL STEP

Mixing Console: Yamaha DMC1000

Additional Hardware/Software: BBC custom made DAC

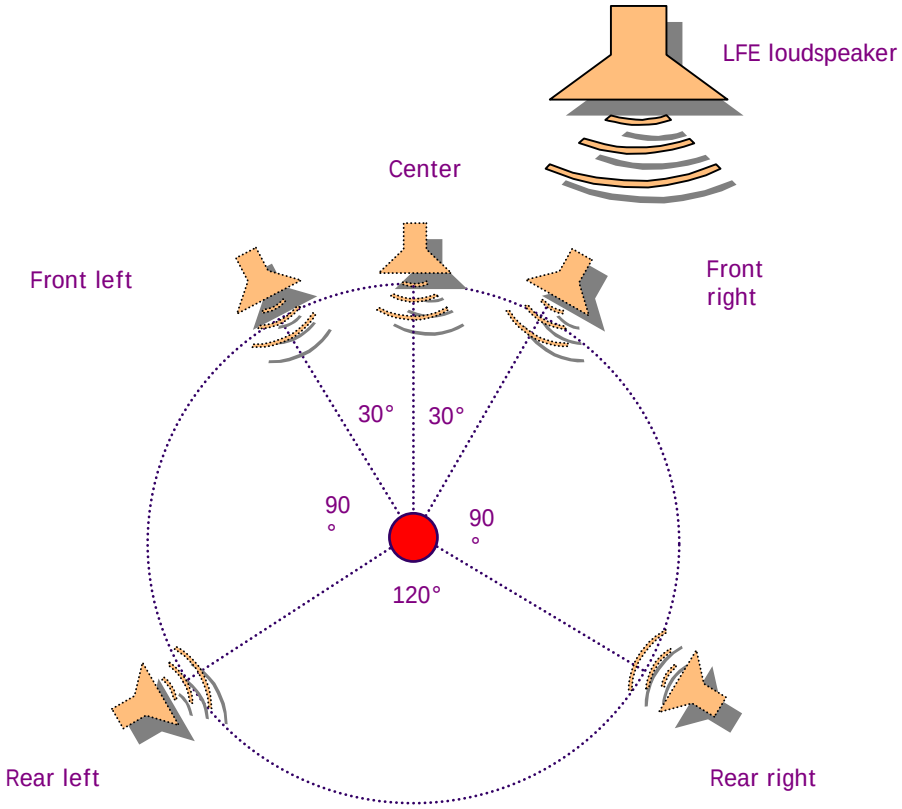
France Telecom - Orange, Rennes, France

Listening room:

Size: Basic shape ...Cube...

Floor area ...20m²...

Height ...2.5m...



Acoustics:

Basic room treatment: n/a

Reverberation time/f: n/a

NC value: n/a

Loudspeakers:

Type (manufacturer): 5x Genelec 8040APM bi-amplified

1x Genelec 7070APM

Frequency & phase response:

http://www.genelec.com/pdf/DS8040a.pdf

http://www.genelec.com/pdf/DS7000_2.pdf

Technical equipment:

Audio workstation:

Hardware, Audio card: PC with a Digigram VX 882 audio board

Software: SEAQ software from CRC Canada, the version dedicated to multichannel audio tests

*Mixing Console: P*reamplifier SPL 2380S

Additional Hardware/Software: an external DAC 24 bit/ 192kHz / 8 channels (Apogee-Rosetta 800).

Institut für Rundfunktechnik (IRT), Munich, Germany

Listening room: meets the Recommendations EBU doc. Tech 3276 including EBU doc.Tech 3276 (Suppl. 1) as well as ITU-R Recommendation BS.1116.

Loudspeakers:

Type (manufacturer): 5x MEG RL922 (Musikelectronic Geithain)

2x Genelec subwoofers (1092A + 1094A)

Technical equipment:

Audio workstation:

Hardware:

Audio card:

Software:

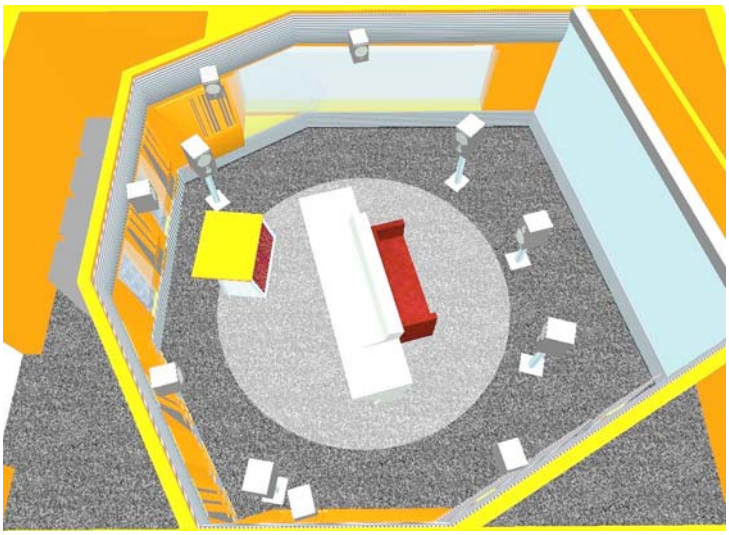
Mixing Console: Yamaha O2R

Additional Hardware/Software: Yamaha DAC

ORF – Austrian Broadcasting Corporation, Vienna, Austria

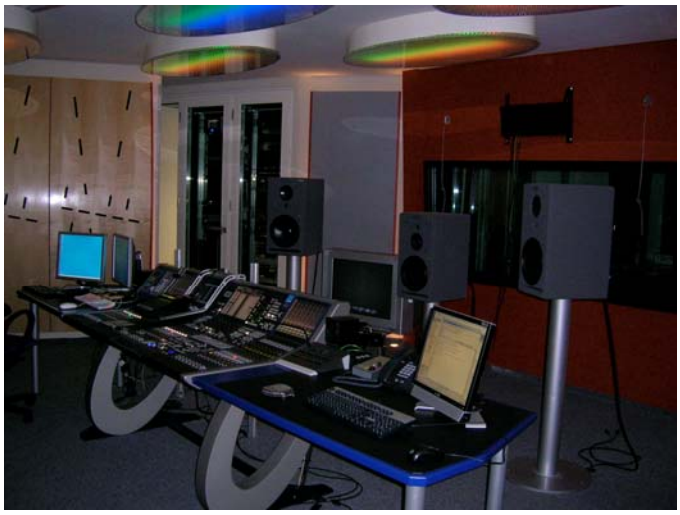
Listening rooms: Post-production studios SK 1 and SK 3

SK1:



SK2:

SK3:



Size:

Basic shape ……asymmetric hexagon

Floor area ……54m^2 (SK 1); 48m^2 (SK 3)

Height ……2.8m

Reference listening position……..according to ITU-R BS.775-1;

Loudspeaker distance: 3.5m (SK 1)

2.85m (SK 3)

placement according to ITU-R BS.775-1

Acoustics:

Basic room treatment: Reflection panels, diffusion elements, Helmholtz resonators

Reverberation time/f: 0.22s av. (SK 3)

NR (Noise rating) value: 25

Loudspeakers:

Type (manufacturer): Studer A5 active monitors

Audio workstation:

Hardware, Audio card: Customized PC with Windows XP

RME Hammerfall MADI card

Software: Nuendo 3.0

Mixing Console: Lawo mc2 82 (SK 1), Lawo mc2 66 (SK 3)

Additional Hardware/Software: Trinnov Optimizer (SK 3)

Radio France, Paris, France

Listening room:

Size:

Basic shape Old natural echo room / near rectangular

Floor area 41m^2…(9.4 Lo ; 4.6 La)……

Height about 4.8m……

Volume : about 200m^3

Reference listening position…about 2.2m almost following ITU BS. 775-2

Acoustics:

Basic room treatment: diffusers, absorbers, panel absorbers
Reverberation time/f 0.3
NC value first round 25; second 30

Loudspeakers:

Type (manufacturer) …Genelec 1032

Technical equipment:

Audio workstation:

Hardware, Audio card …P4, Lynx Two

Software Wavelab 6; Lynx Two mixer

RAI CRIT – Centro Ricerche Innovazione Tecnologica, Turin, Italy

Listening room:

Size:

Basic shape rectangular

Floor area: WxD 500 x 800 cm

Height: 335 cm

Reference listening position: centre of the ring beam 210 cm

Acoustics:

Basic room treatment: drapery, diaphragmatic absorber

Reverberation time/fn/a

NC value n/a

Loudspeakers:

Type (manufacturer) **GENELEC 8050A monitors + 7070A subwoofer**
Frequency/phase response:

- http://www.genelec.com/pdf/DS8050a.pdf
- http://www.genelec.com/pdf/DS7000_2.pdf

Technical equipment:

Audio workstation:

Hardware, Audio card: Windows XP Workstation with RME9652 digital interface

Software: STEP software (version 1.00) from audio research labs

Mixing Console: Yamaha DM2000

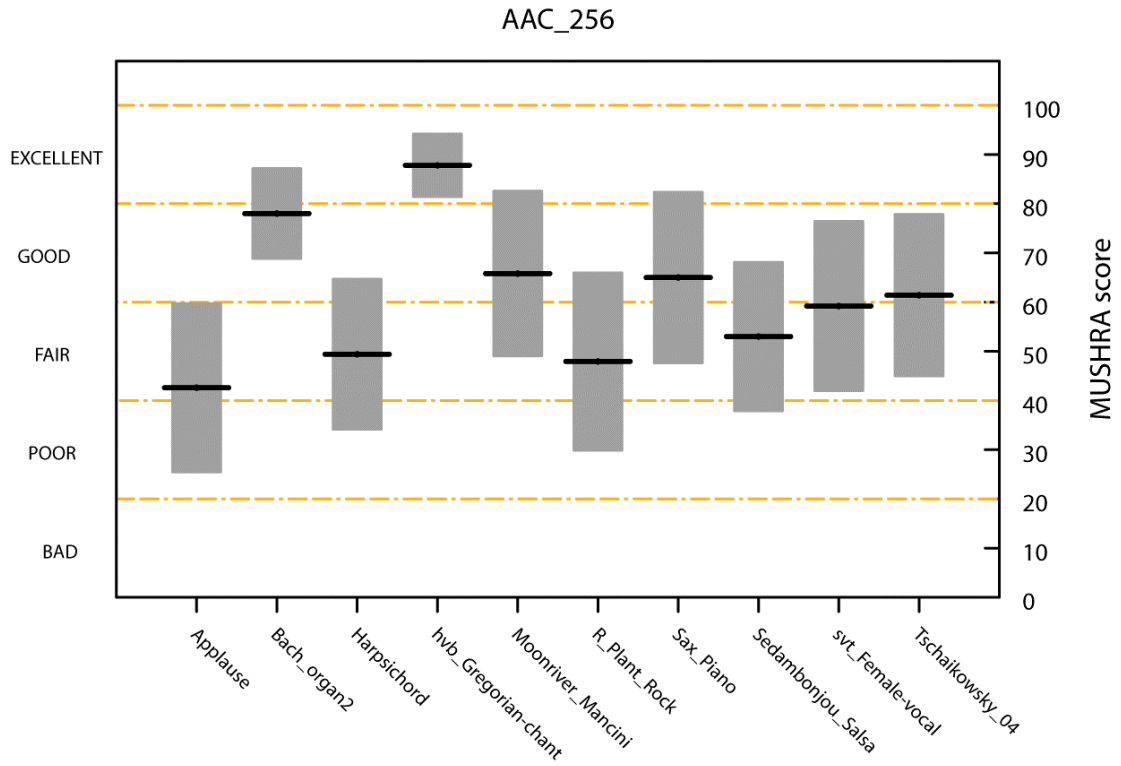
Page intentionally left blank. This document is paginated for recto-verso printing

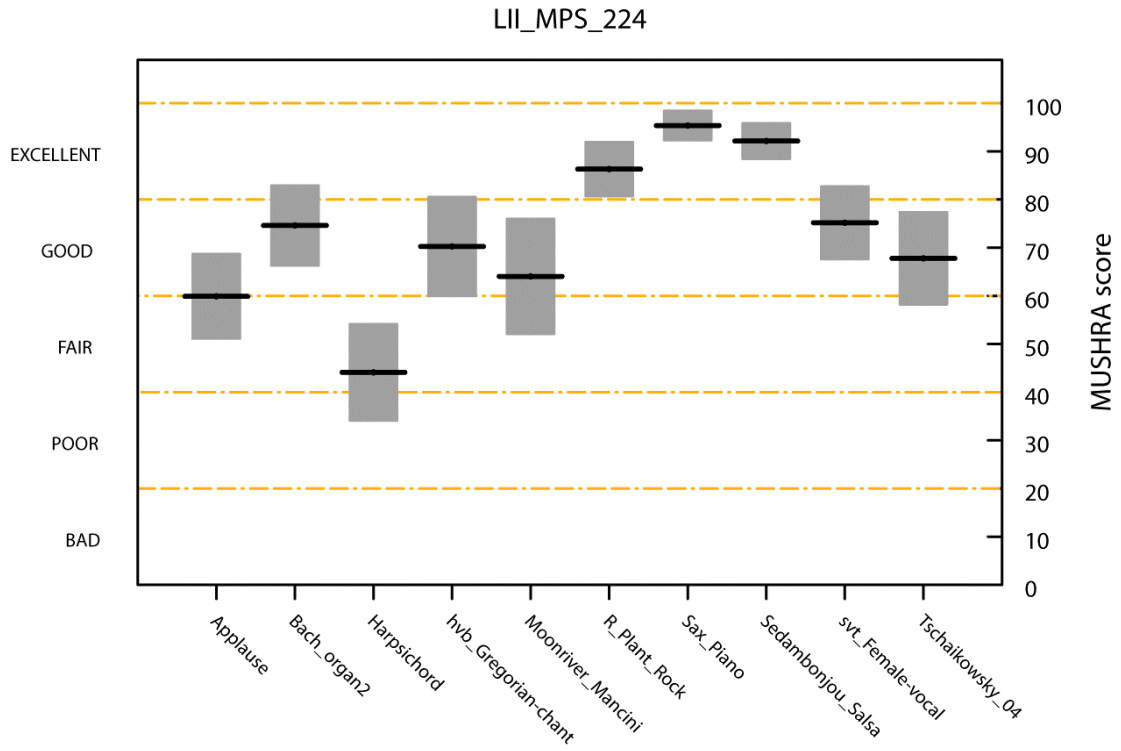
Appendix 4: Detailed test results

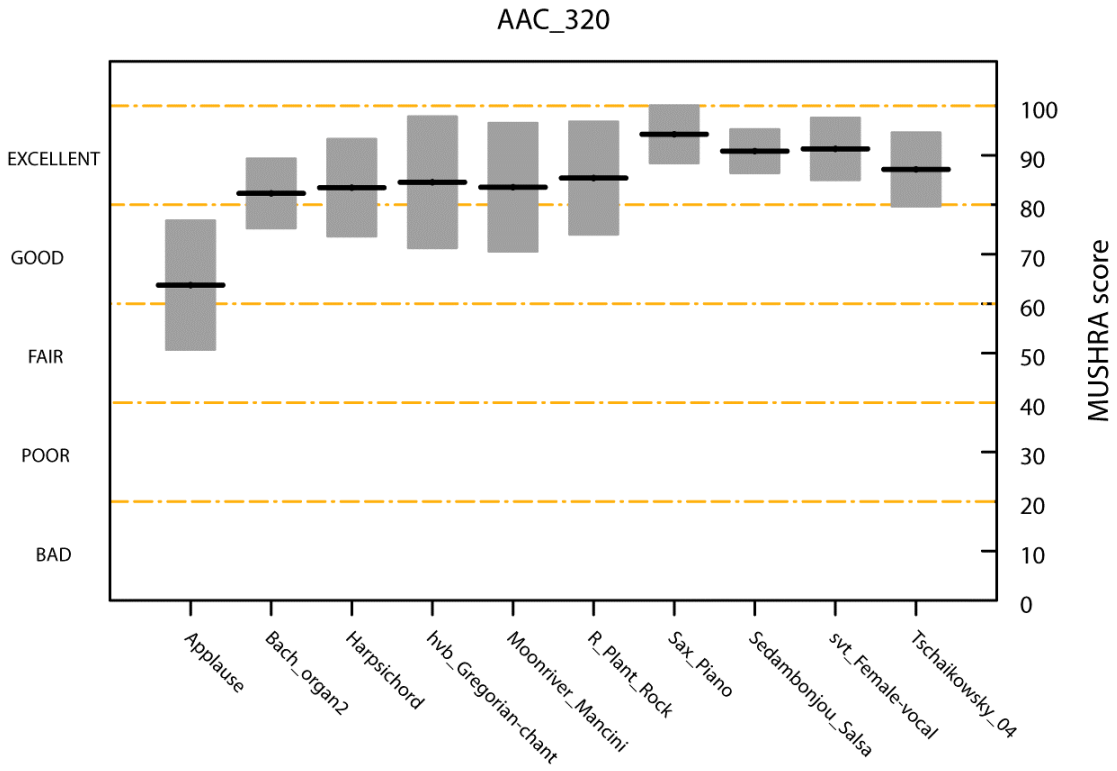
Phase 1

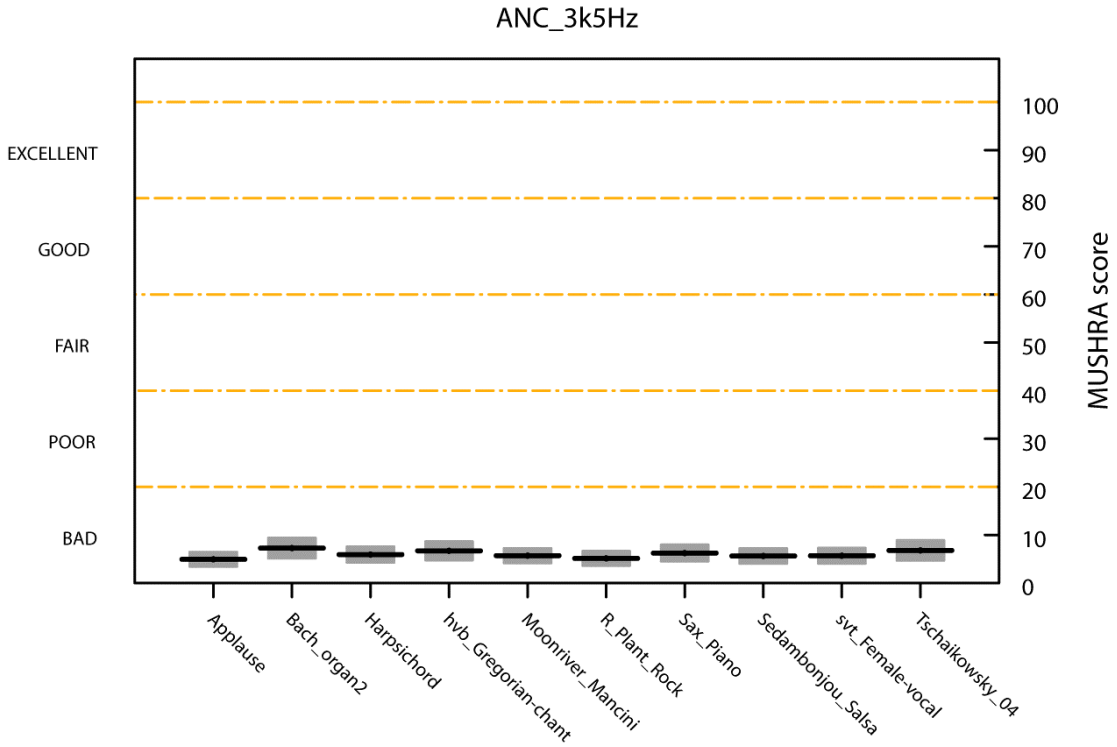
Averages for each codec over all test items - Phase 1

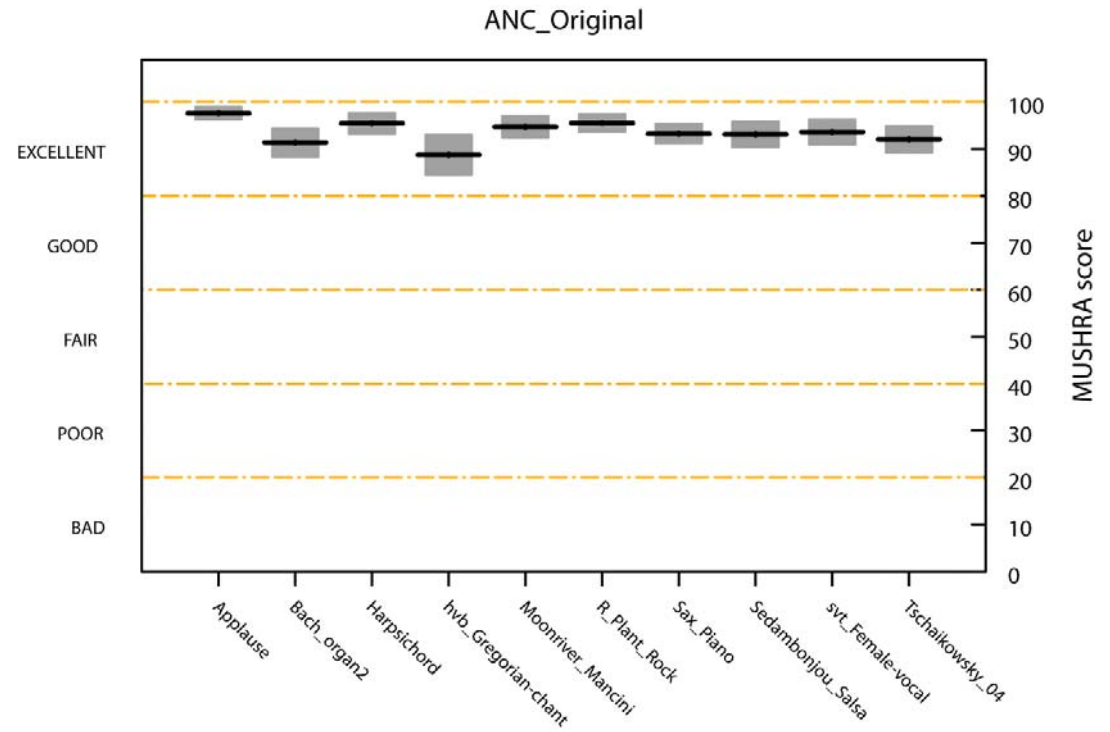
These graphs show how each codec scored for each of the test items. They are useful to highlight any particular strengths or weaknesses with particular types of material for each codec.

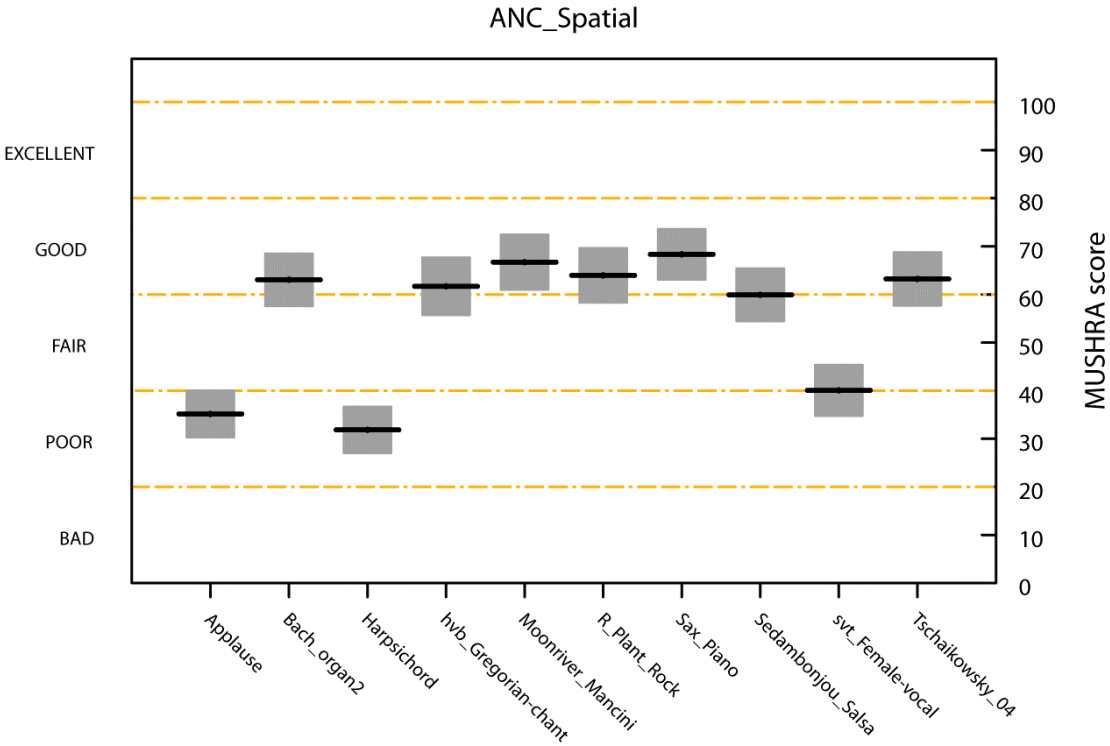


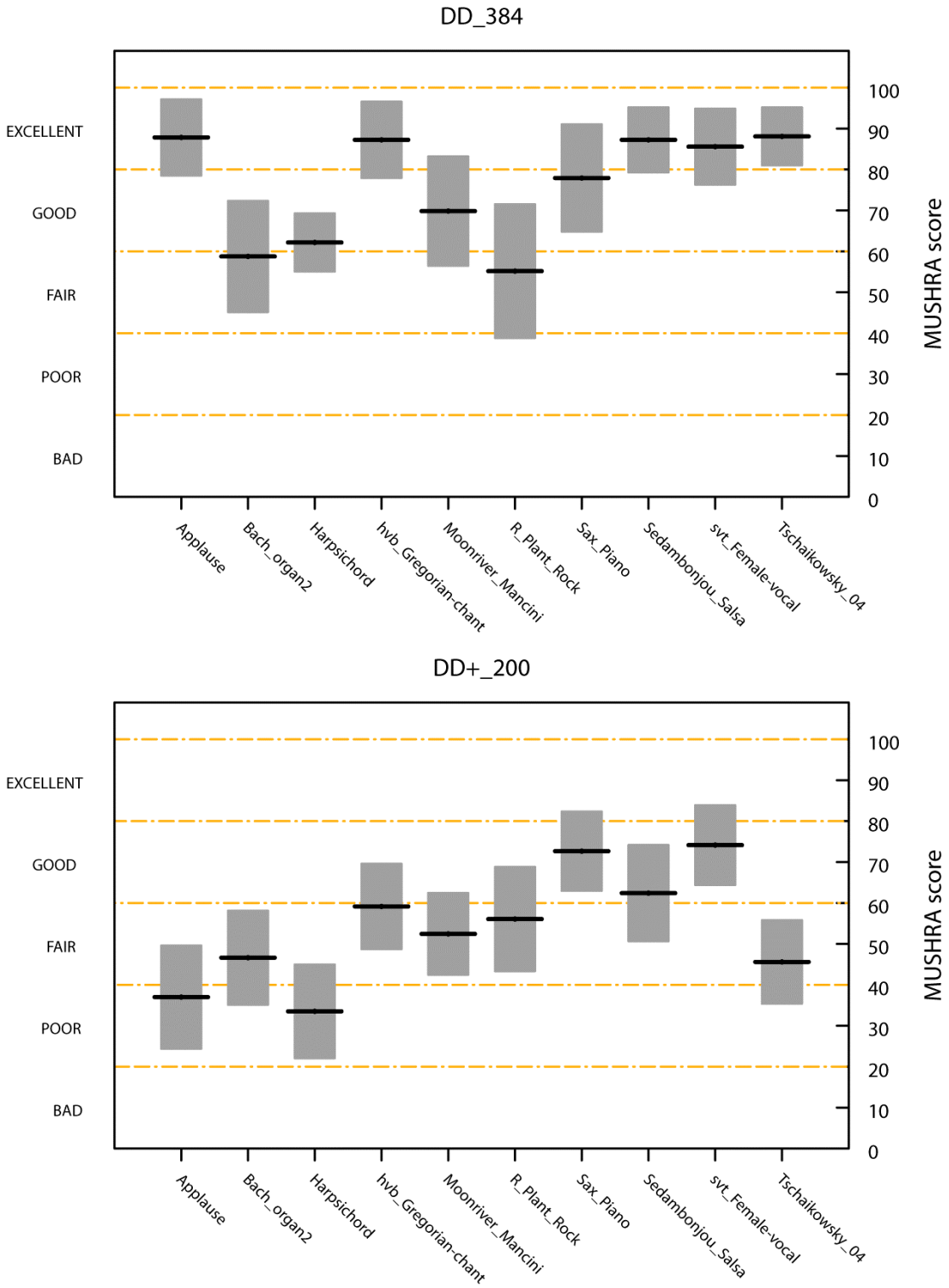


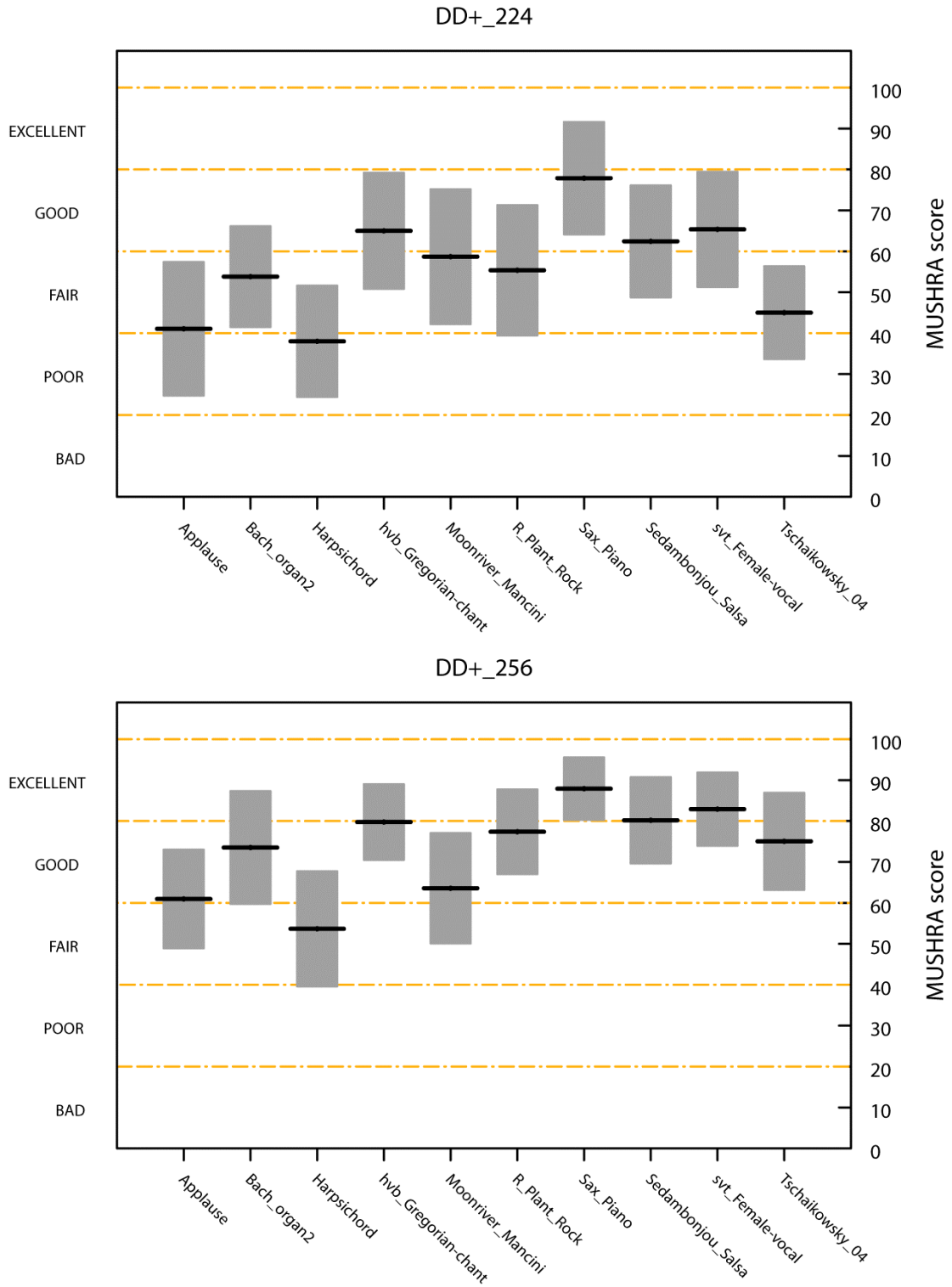

AAC_320
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04

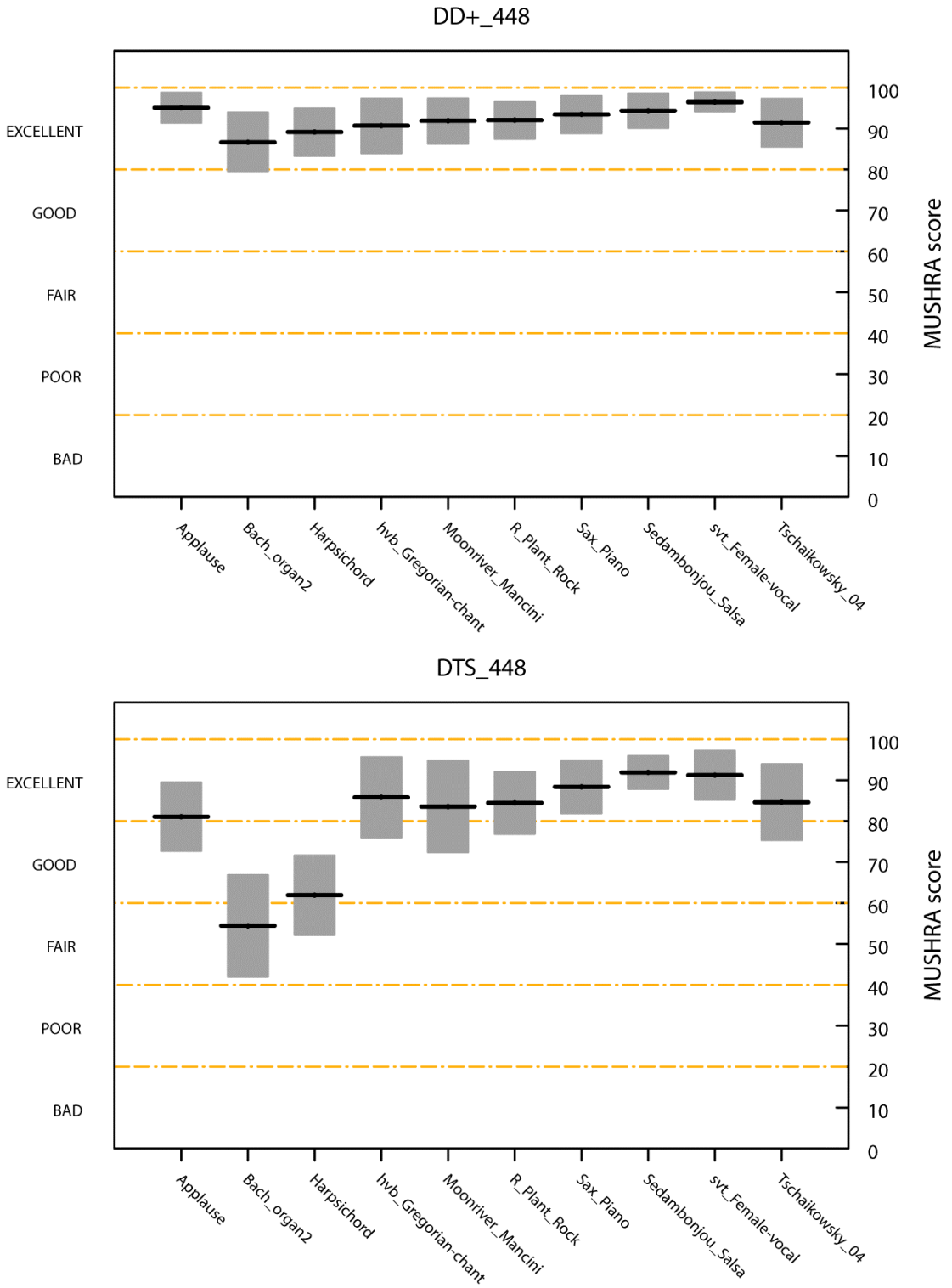

ANC_3k5Hz
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04

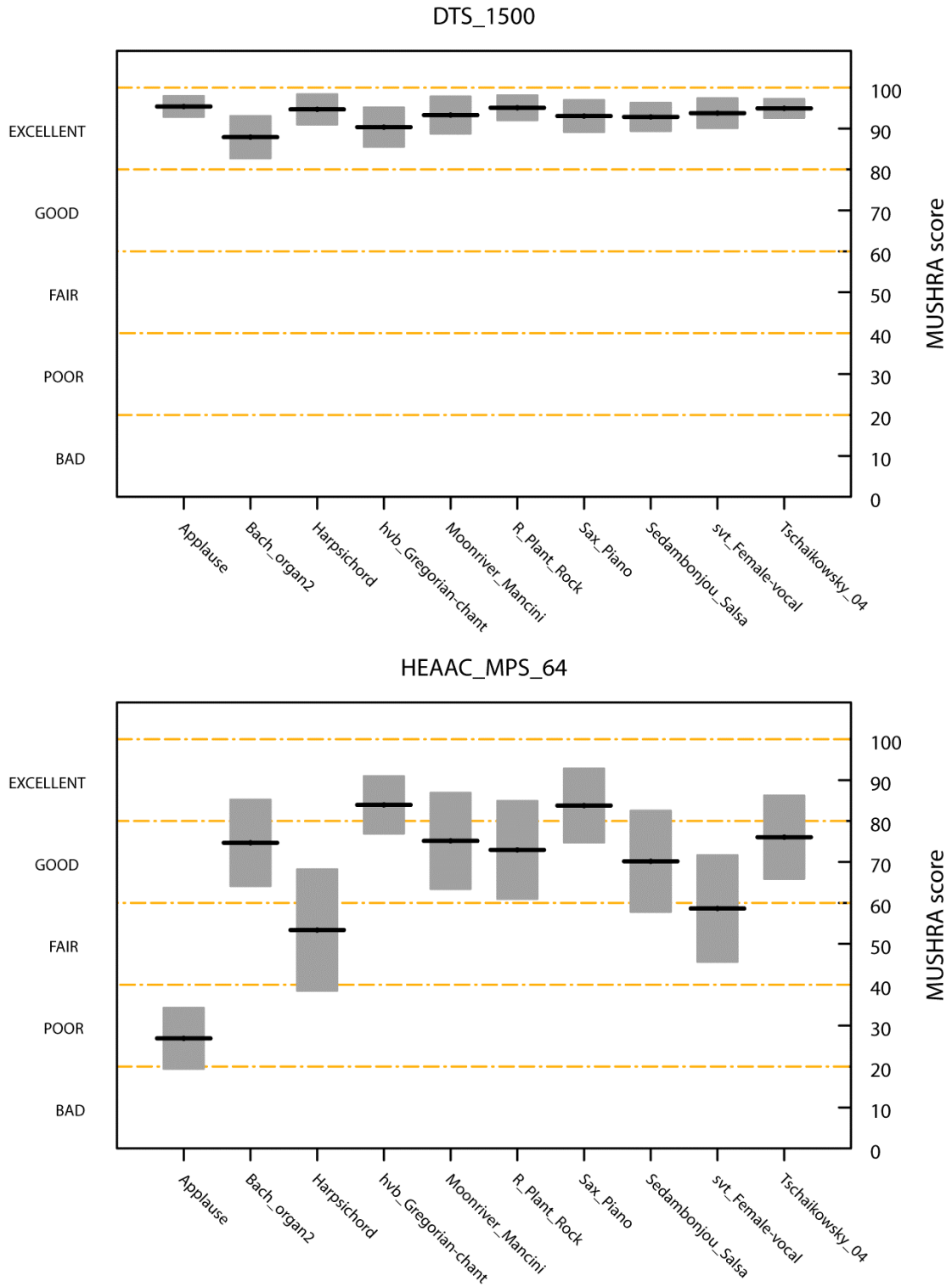

ANC_Original
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04

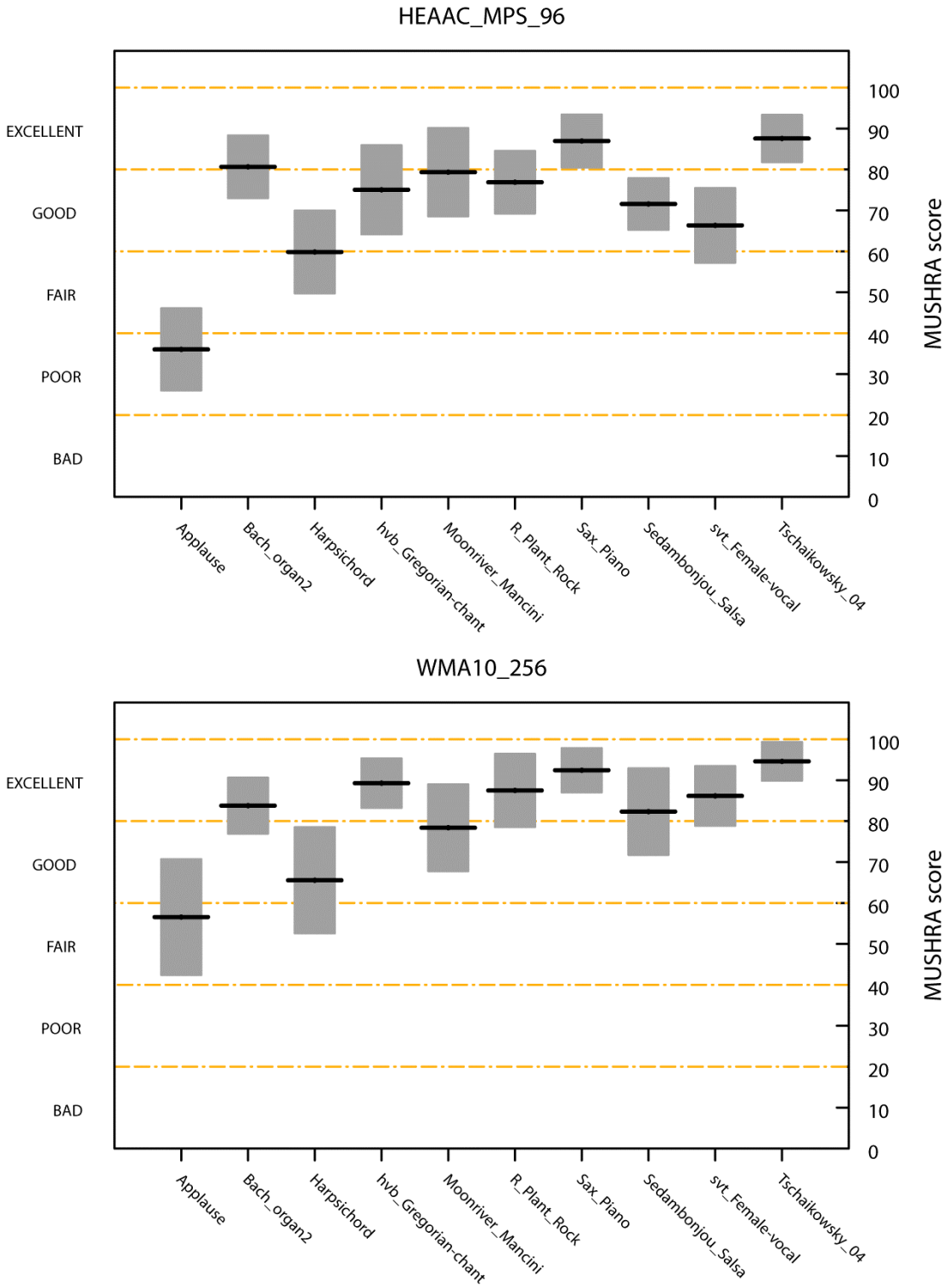

ANC_Spatial
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04

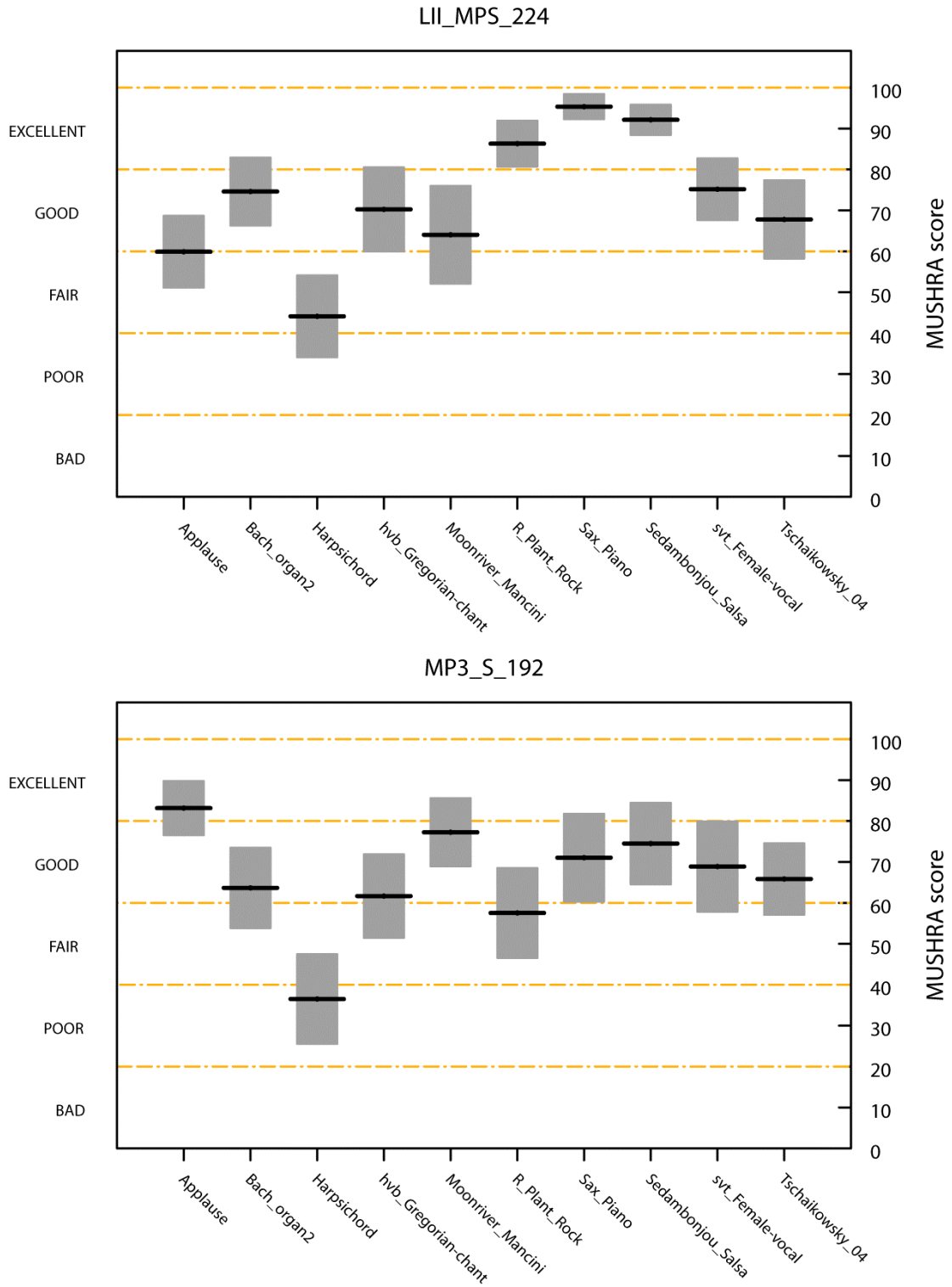

DD_384
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04
DD+_200
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04

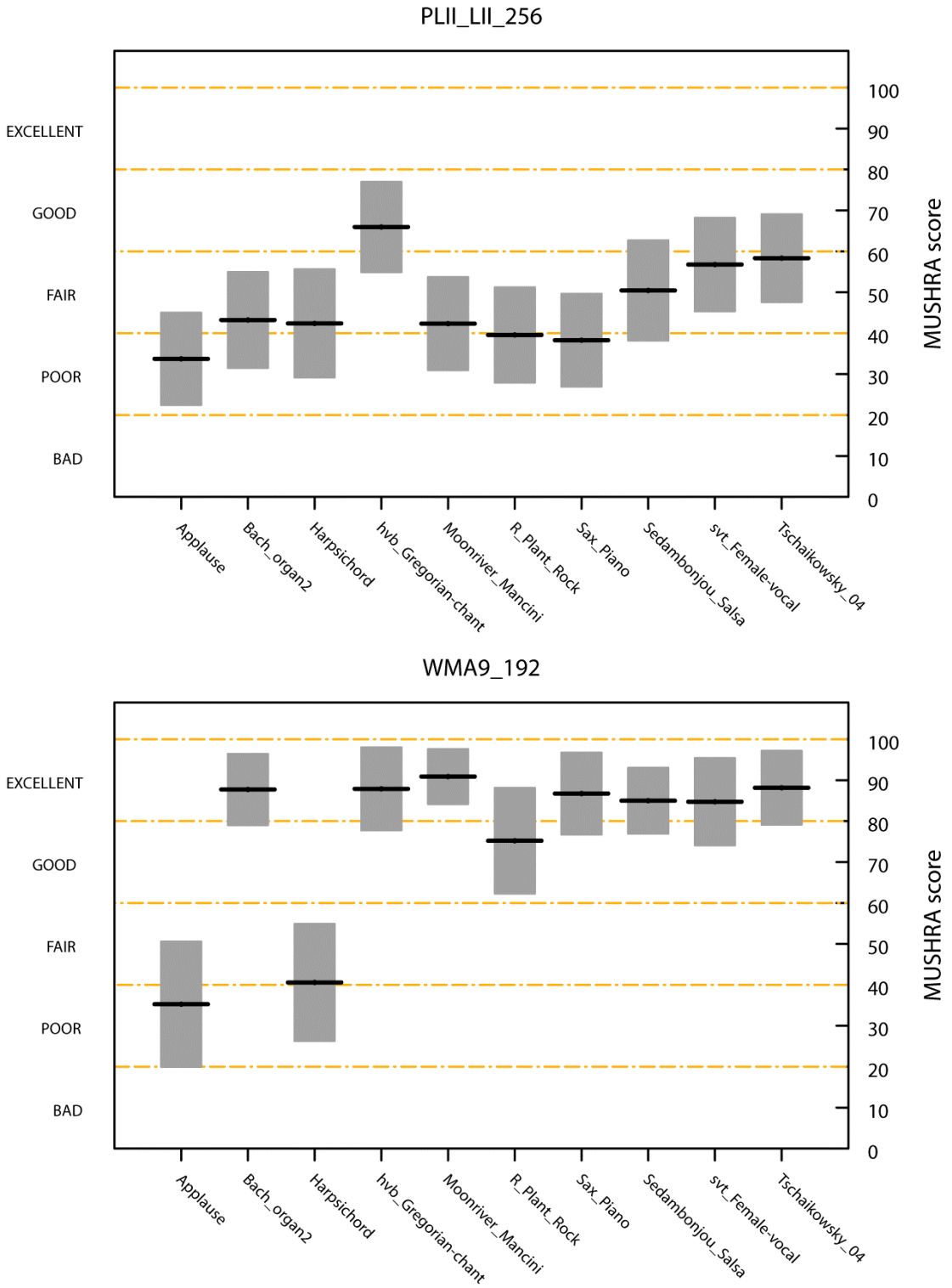

DD+_224
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04
DD+_256
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04

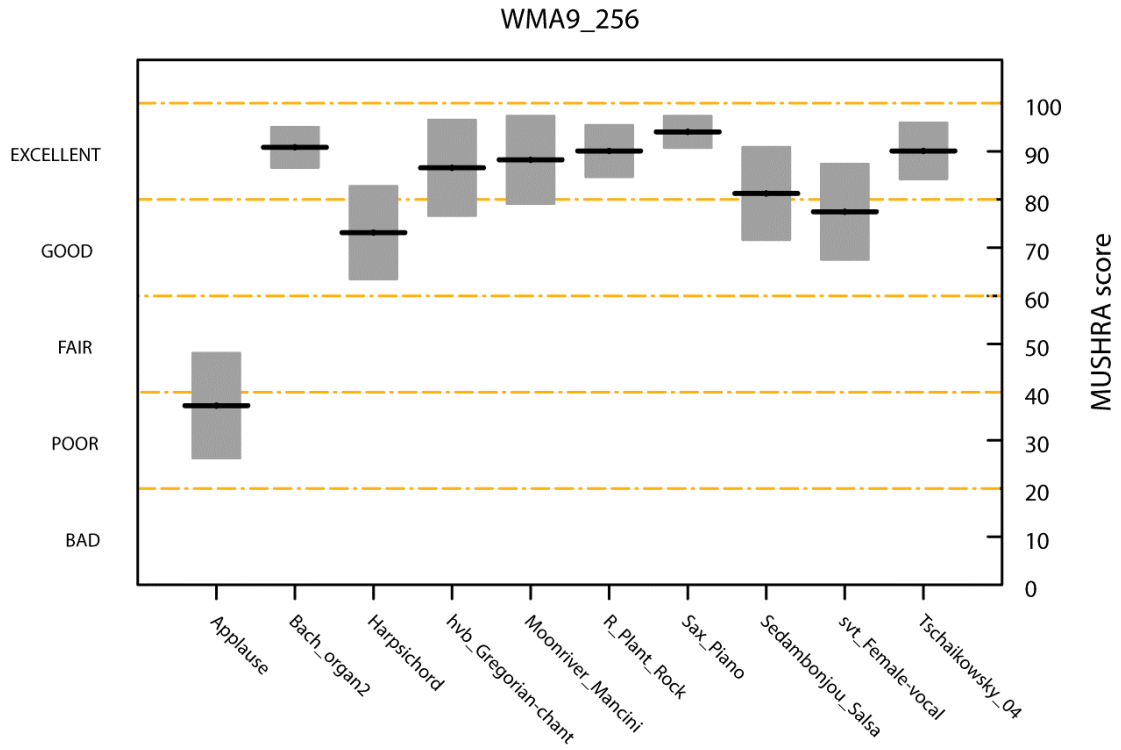

DD+_448
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04
DTS_448
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04

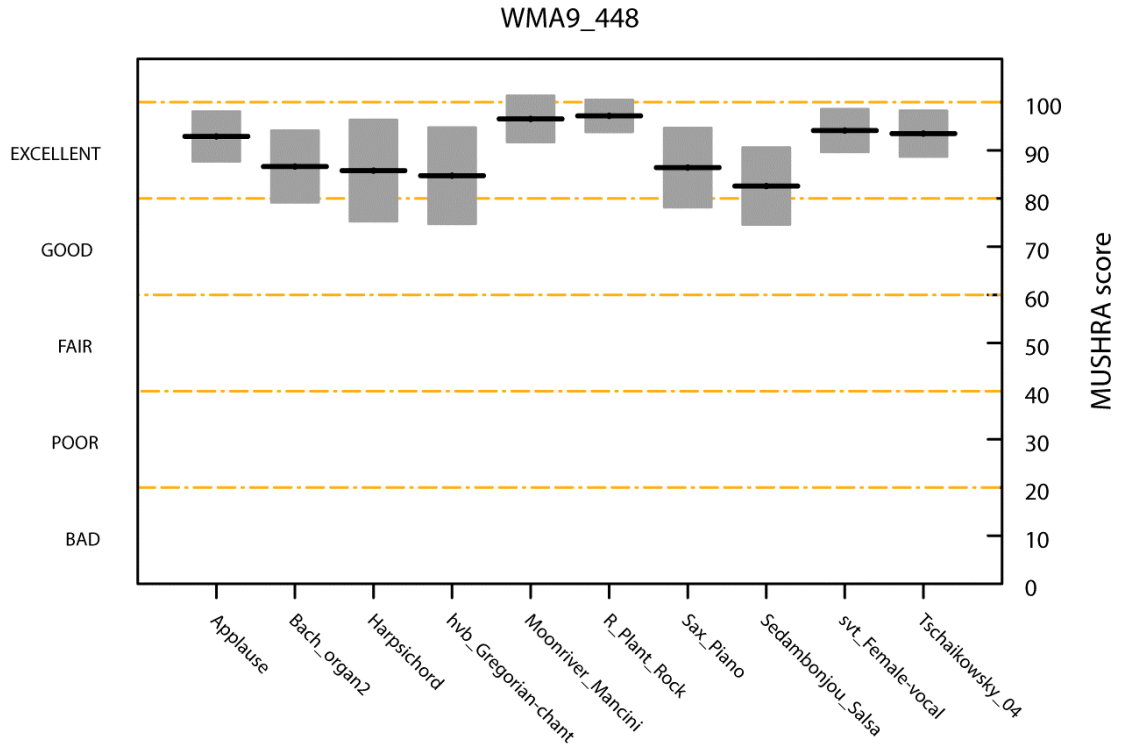

DTS_1500
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04
HEAAC_MPS_64
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04

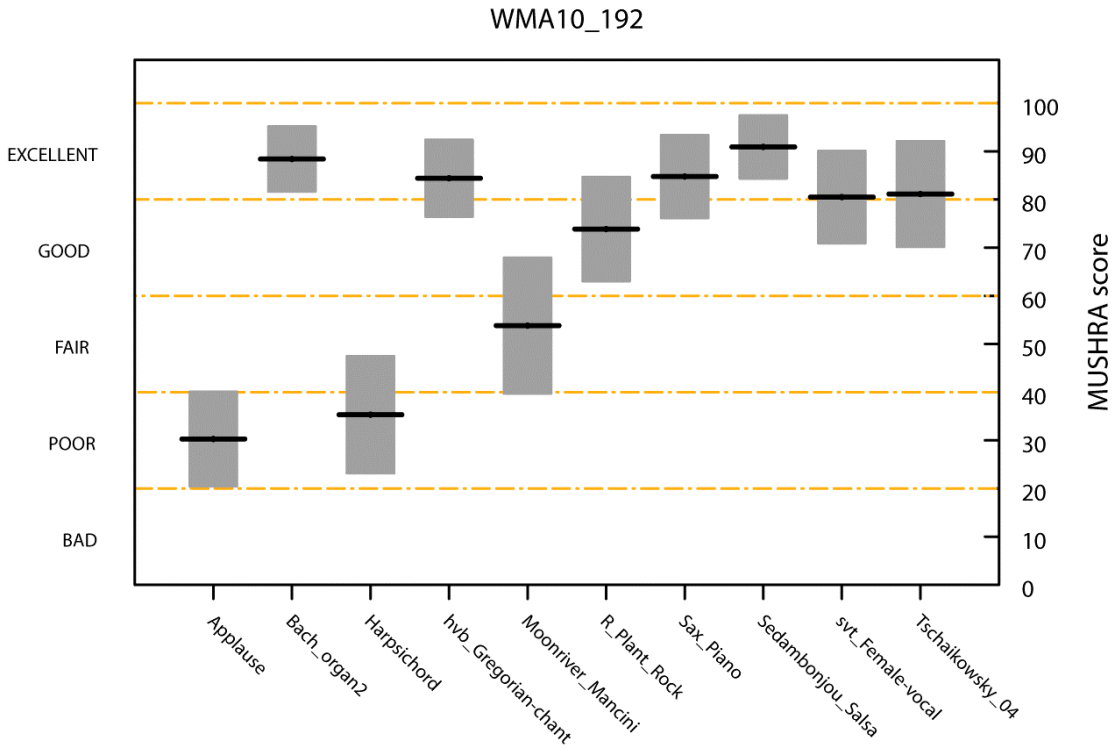

HEAAC_MPS_96
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04
WMA10_256
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04


LII_MPS_224
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04
MP3_S_192
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04


PLII_LII_256
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04
WMA9_192
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04

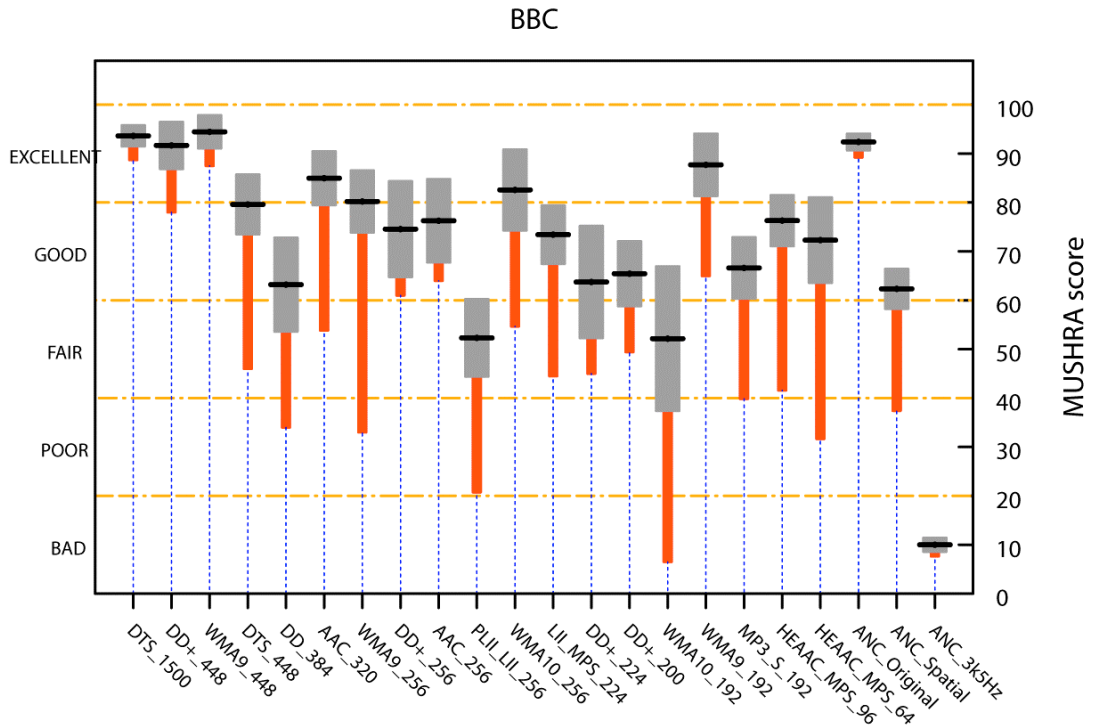

WMA9_256
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04


WMA9_448
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
Applause
Bach_organ2
Harpsichord
hvb_Gregorian-chant
Moonriver_Mancini
R_Plant_Rock
Sax_Piano
Sedambonjou_Salsa
svt_Female-vocal
Tschaikowsky_04

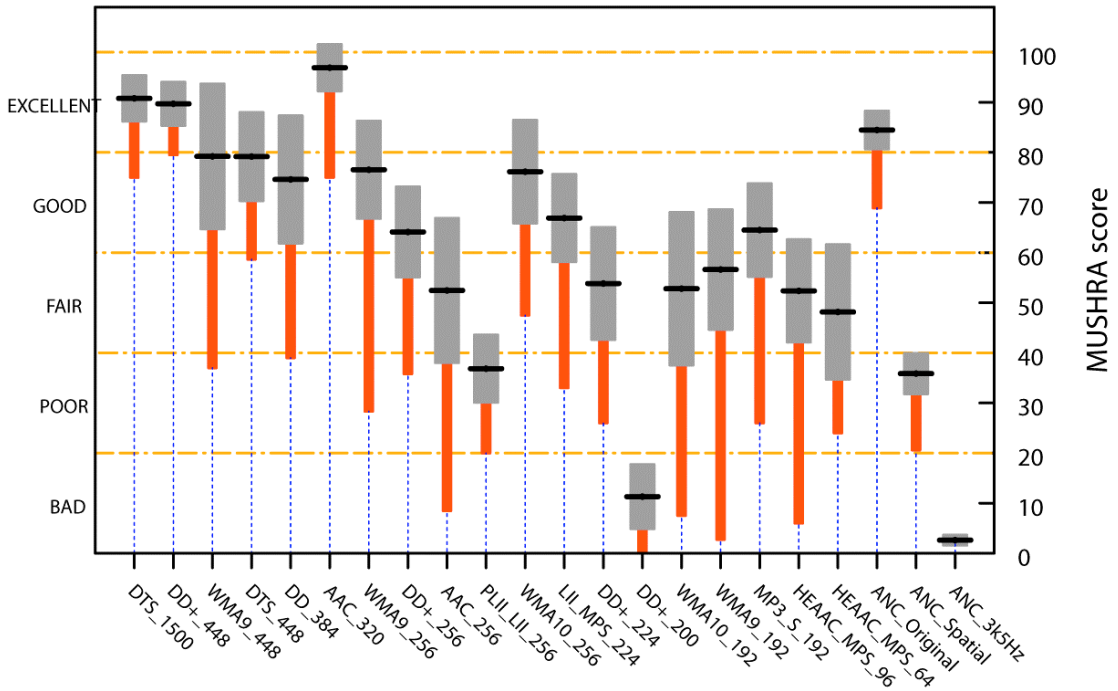


Average scores for each lab over all codecs - Phase 1

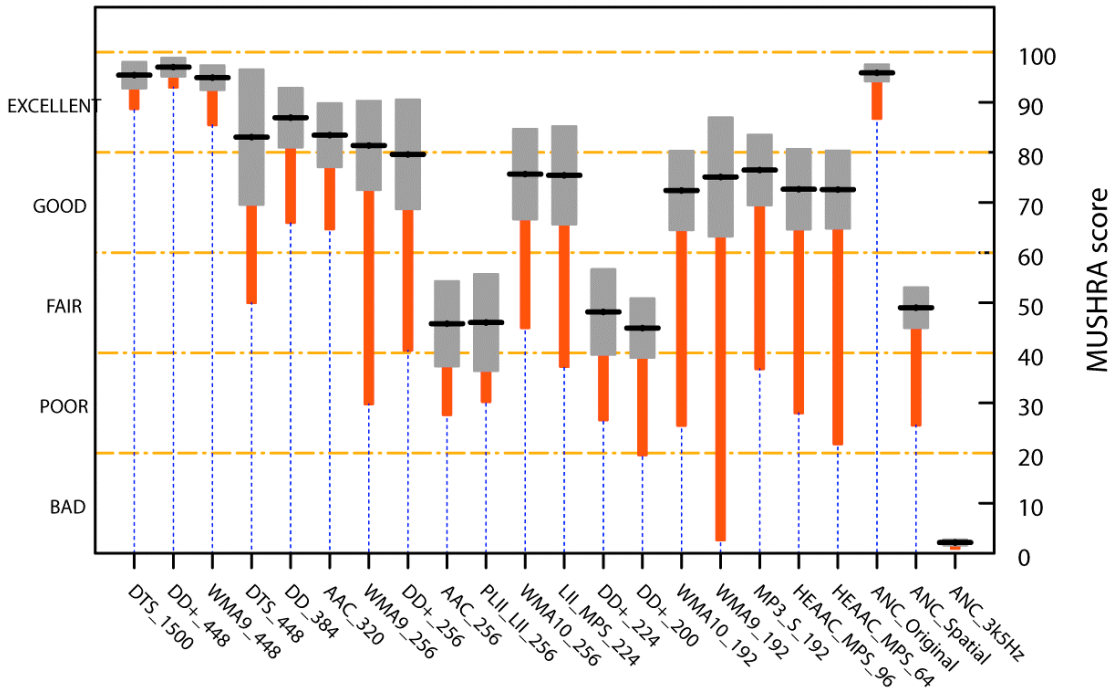
These graphs give the average and worse case scores for each codec on a lab by lab basis. While all the labs tested all the codecs, some listeners were rejected which resulted in some labs not having any scores for some codecs, hence the missing marks on some graphs.

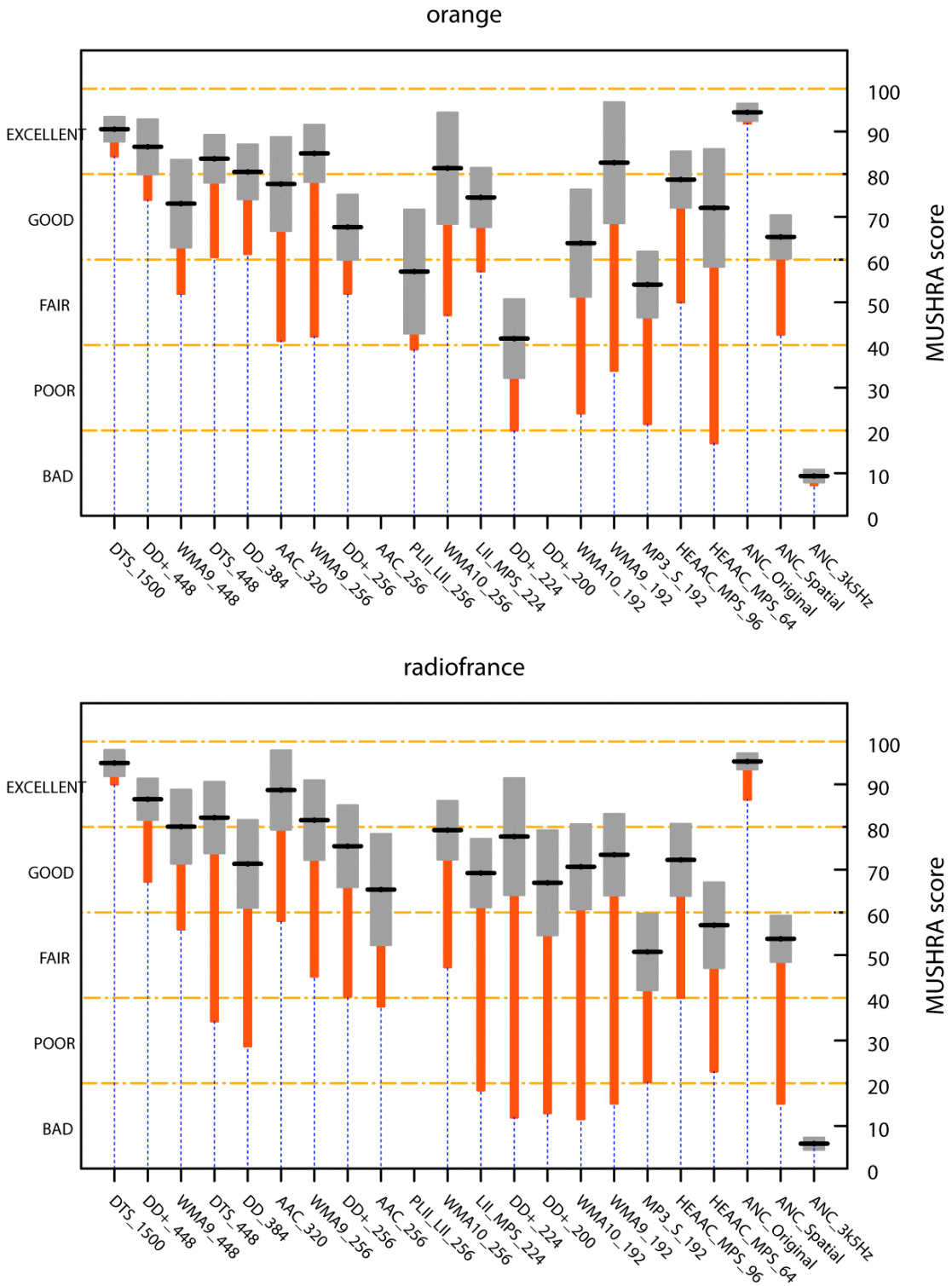


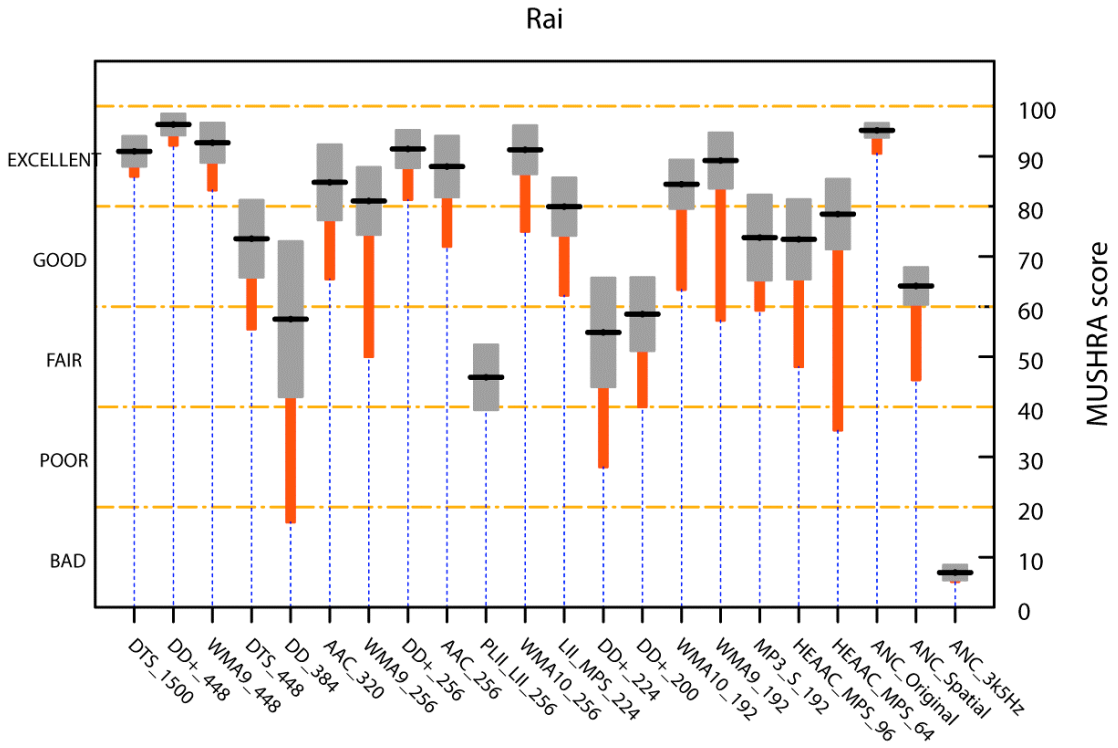
BR


EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
DTS_1500
DD+_448
WMA9_448
DTS_448
DD_384
AAC_320
WMA9_256
DD+_256
AAC_256
PLII_LII_256
WMA10_256
LII_MPS_224
DD+_224
DD+_200
WMA10_192
WMA9_192
MP3_S_192
HEAAC_MPS_96
HEAAC_MPS_64
ANC_Original
ANC_Spatial
ANC_3k5Hz

IRT


EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
DTS_1500
DD+_448
WMA9_448
DTS_448
DD_384
AAC_320
WMA9_256
DD+_256
AAC_256
PLII_LII_256
WMA10_256
LII_MPS_224
DD+_224
DD+_200
WMA10_192
WMA9_192
MP3_S_192
HEAAC_MPS_96
HEAAC_MPS_64
ANC_Original
ANC_Spatial
ANC_3k5Hz

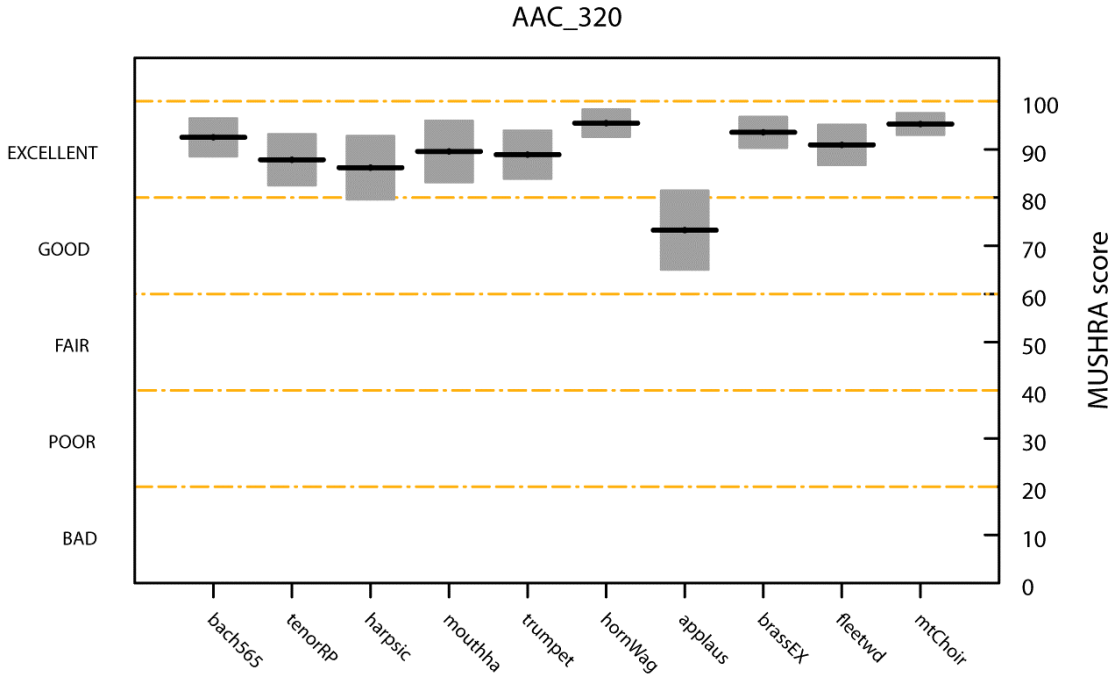

orange
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
DTS_1500
DD+_448
WMA9_448
DTS_448
DD_384
AAC_320
WMA9_256
DD+_256
AAC_256
PLII_LII_256
WMA10_256
LII_MPS_224
DD+_224
DD+_200
WMA10_192
WMA9_192
MP3_S_192
HEAAC_MPS_96
HEAAC_MPS_64
ANC_Original
ANC_Spatial
ANC_3k5Hz
radiofrance
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
DTS_1500
DD+_448
WMA9_448
DTS_448
DD_384
AAC_320
WMA9_256
DD+_256
AAC_256
PLII_LII_256
WMA10_256
LII_MPS_224
DD+_224
DD+_200
WMA10_192
WMA9_192
MP3_S_192
HEAAC_MPS_96
HEAAC_MPS_64
ANC_Original
ANC_Spatial
ANC_3k5Hz

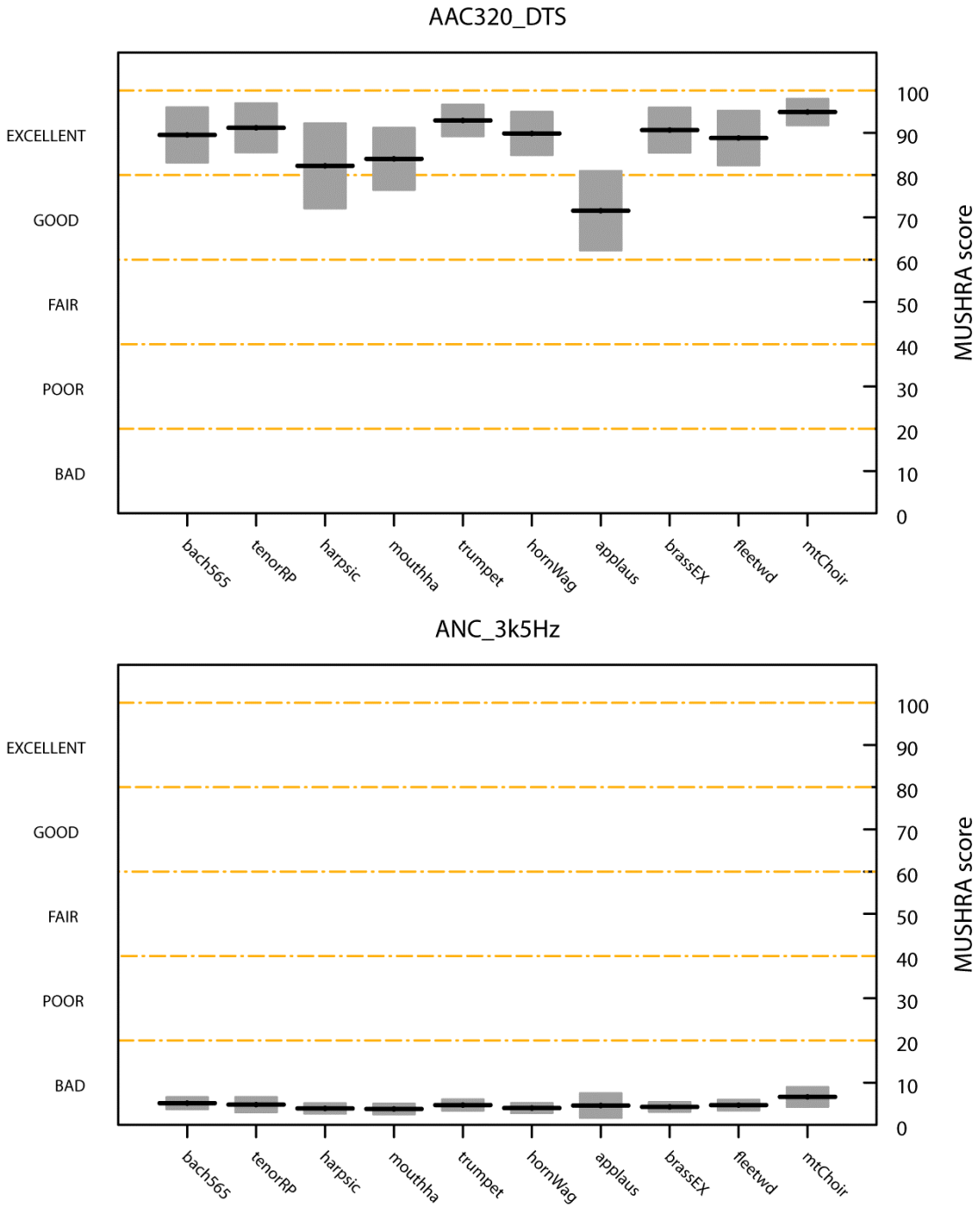


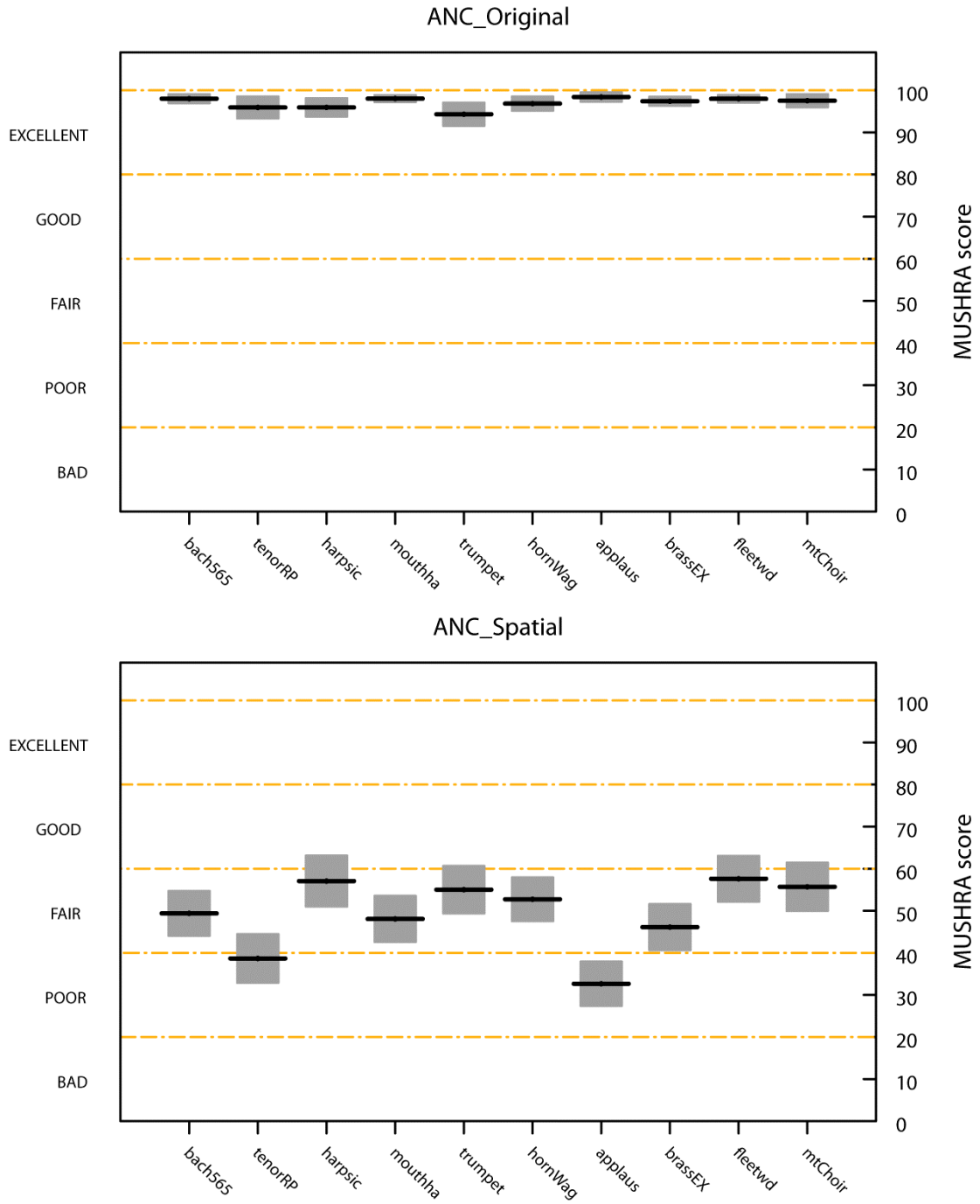
Phase 2

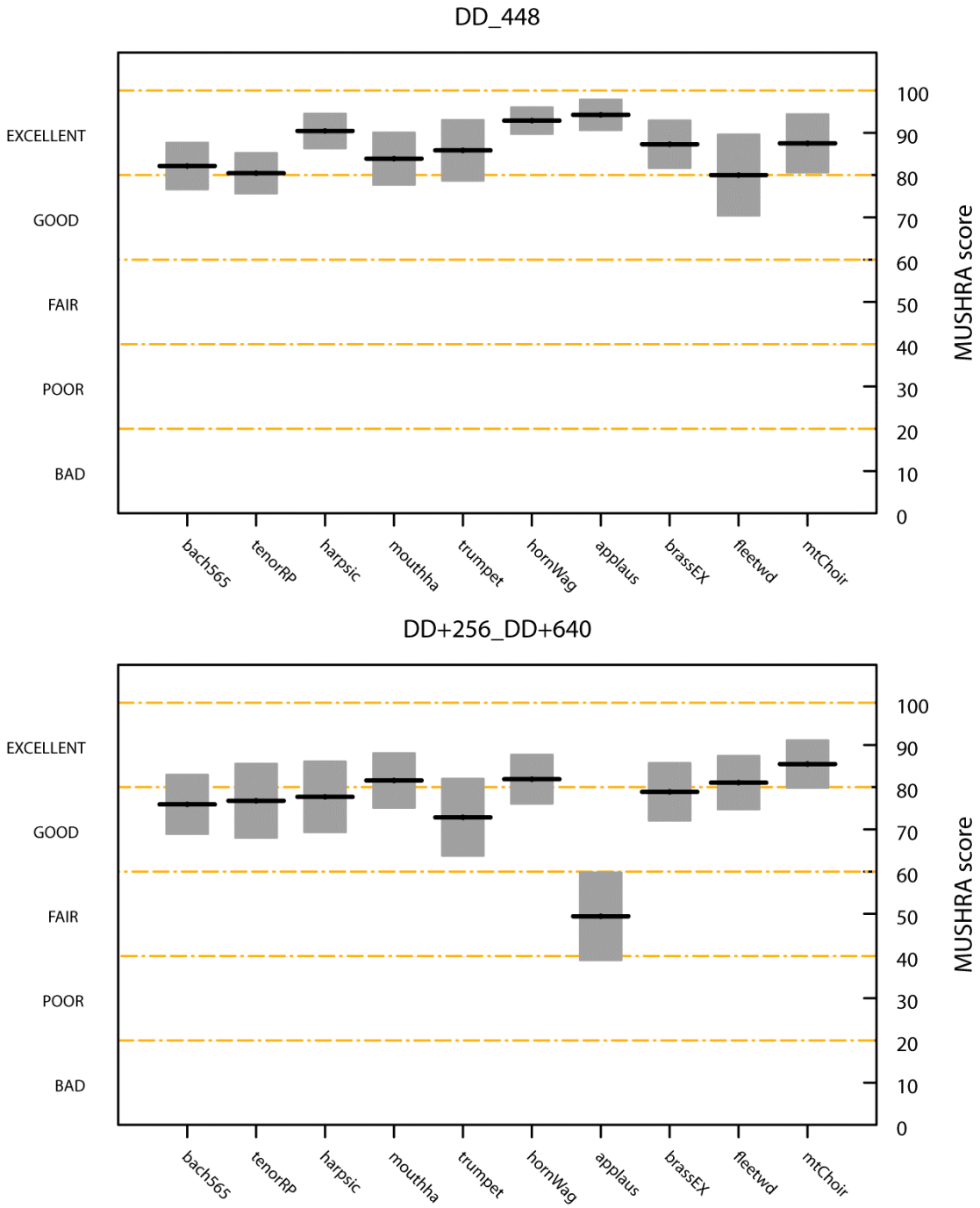
Averages for each codec over all test items -Phase 2

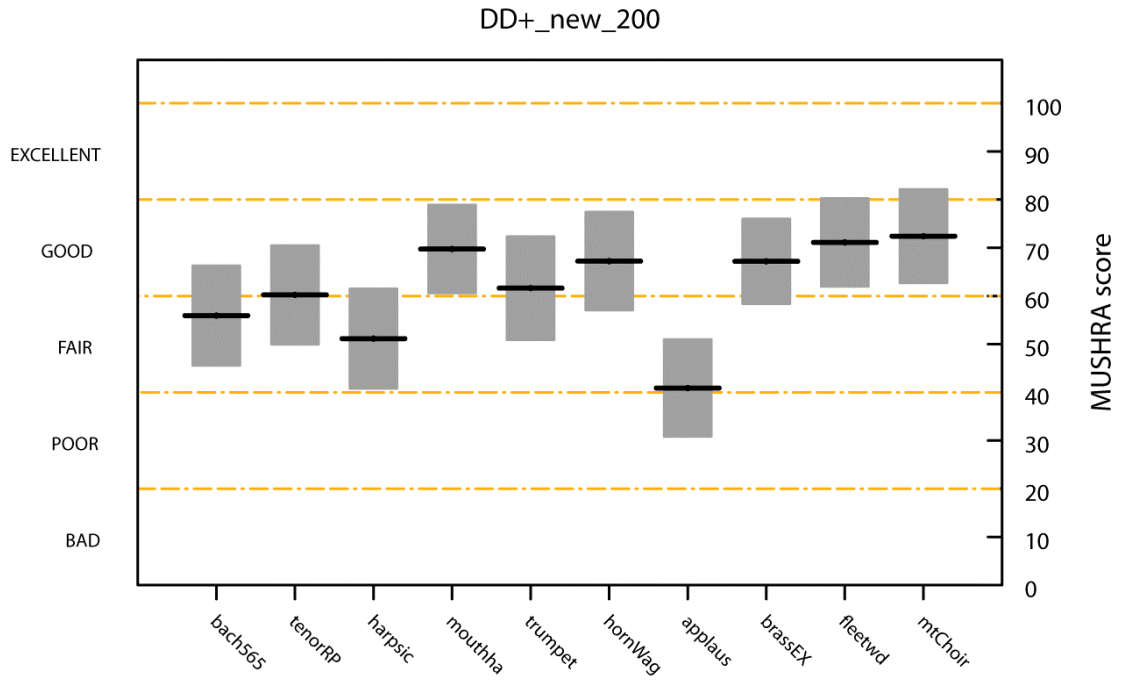
These graphs show how each codec scored for each of the test items. They are useful to highlight any particular strengths or weaknesses with particular types of material for each codec.

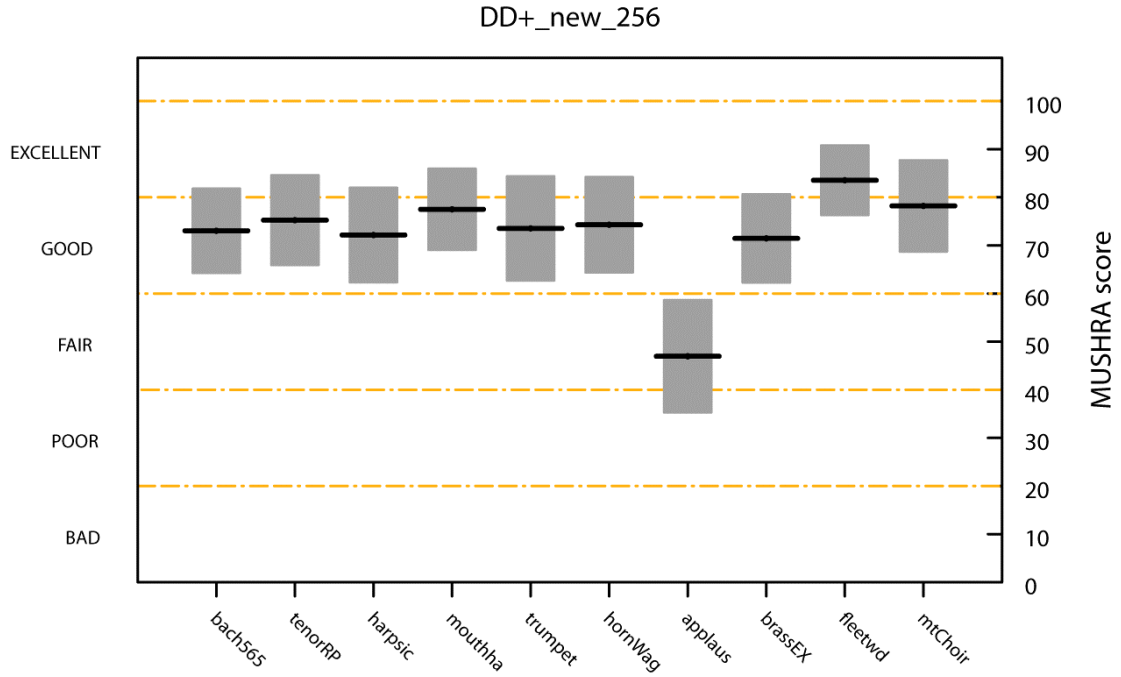


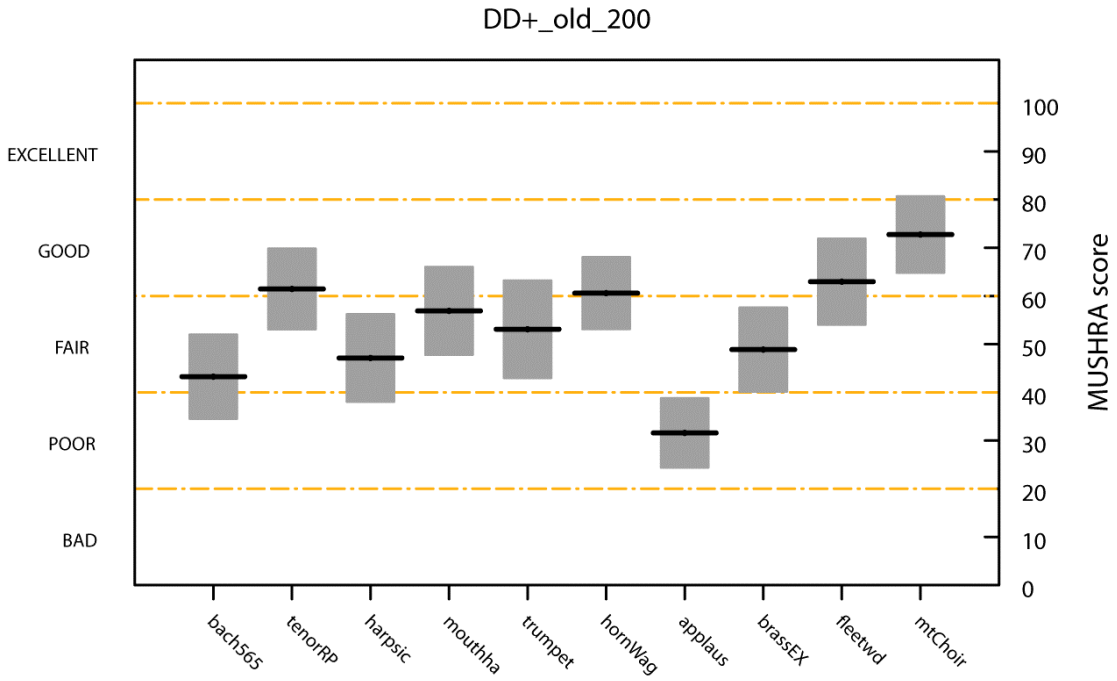

AAC320_DTS
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir
ANC_3k5Hz
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir

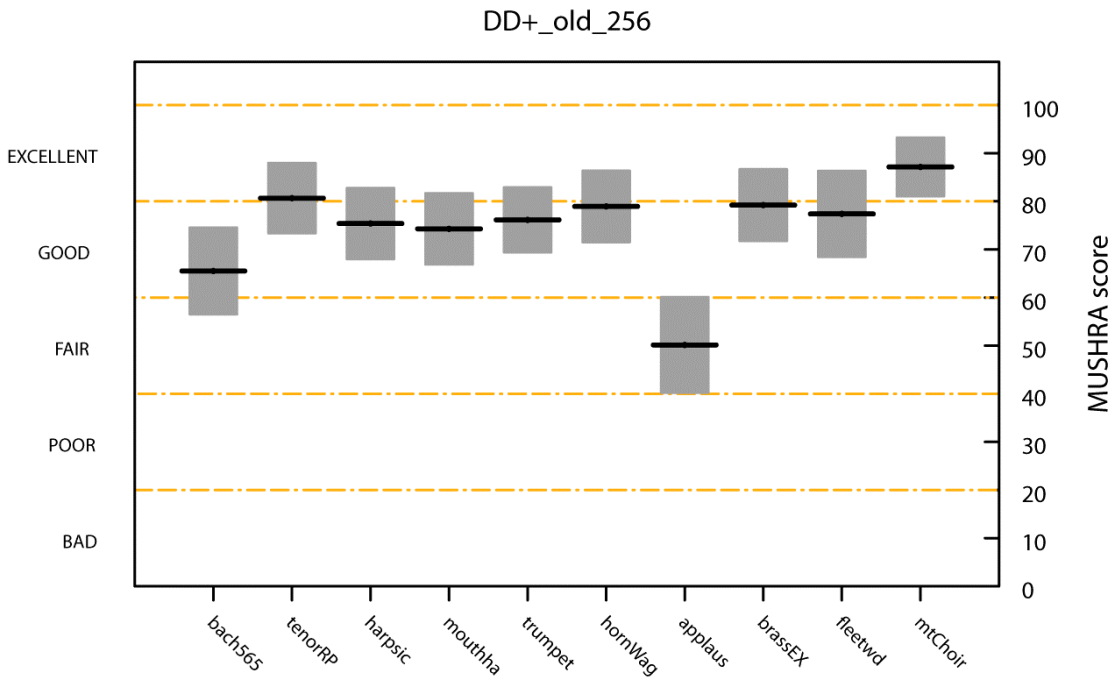

ANC_Original
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir
ANC_Spatial
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir

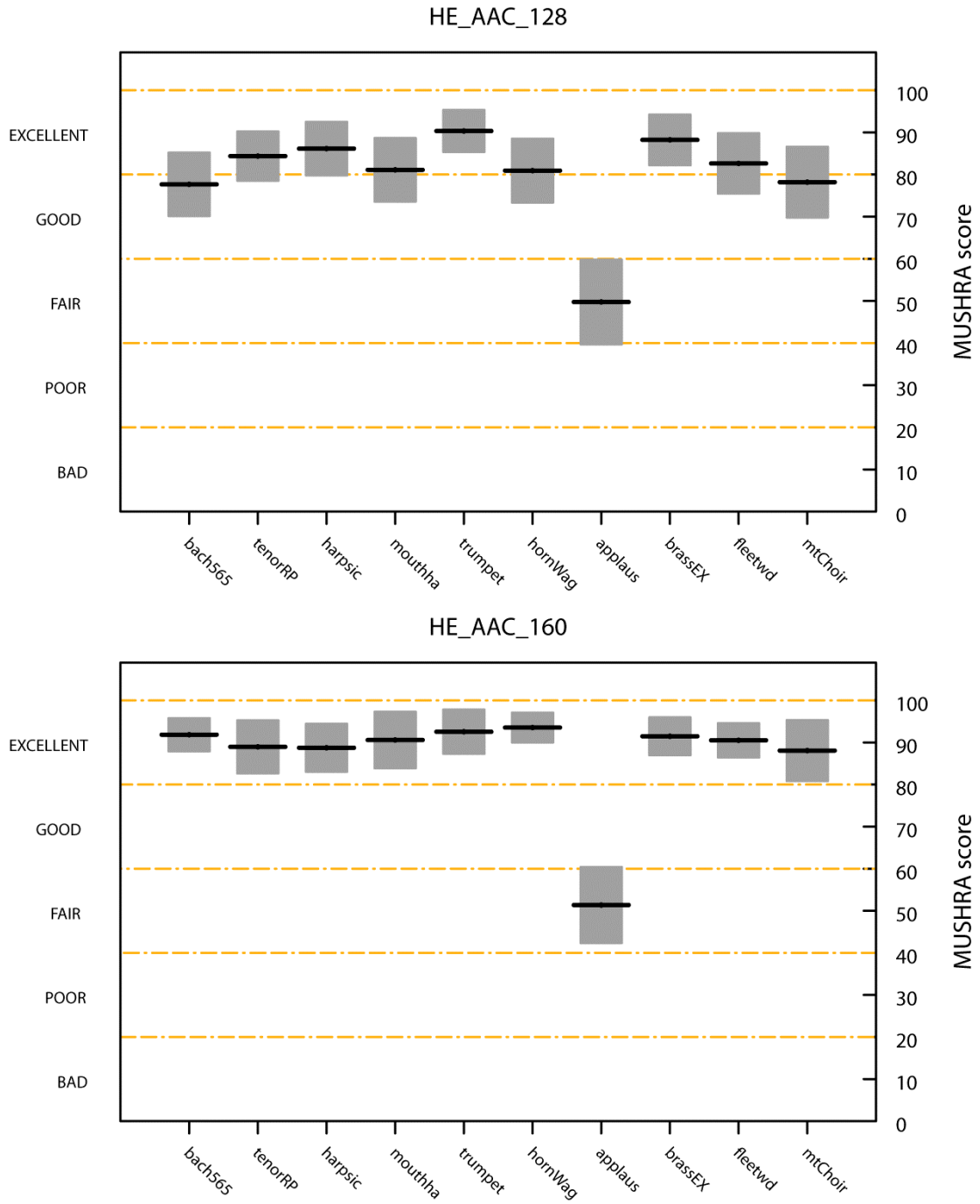

DD_448
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir
DD+256_DD+640
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir

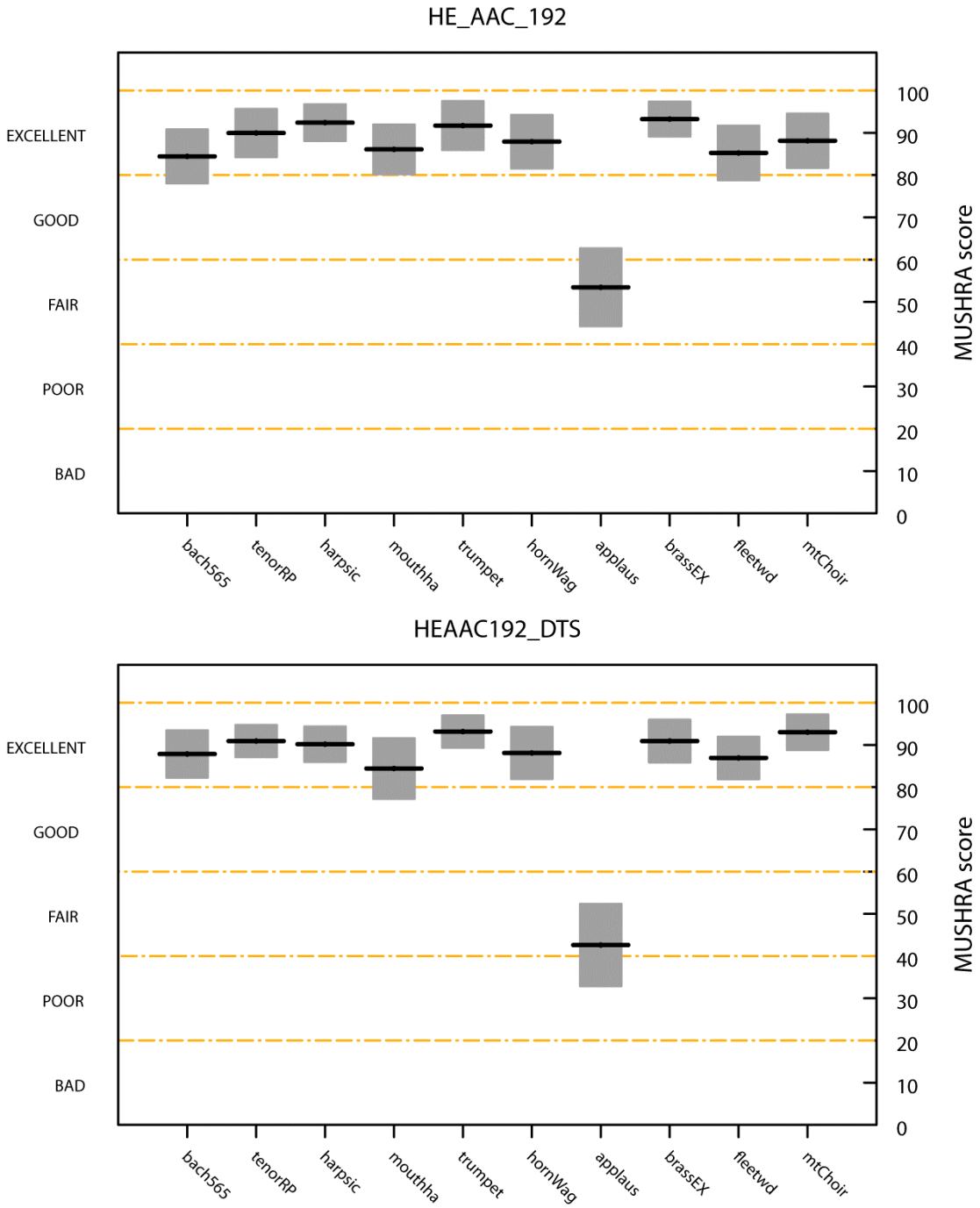

DD+_new_200
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir

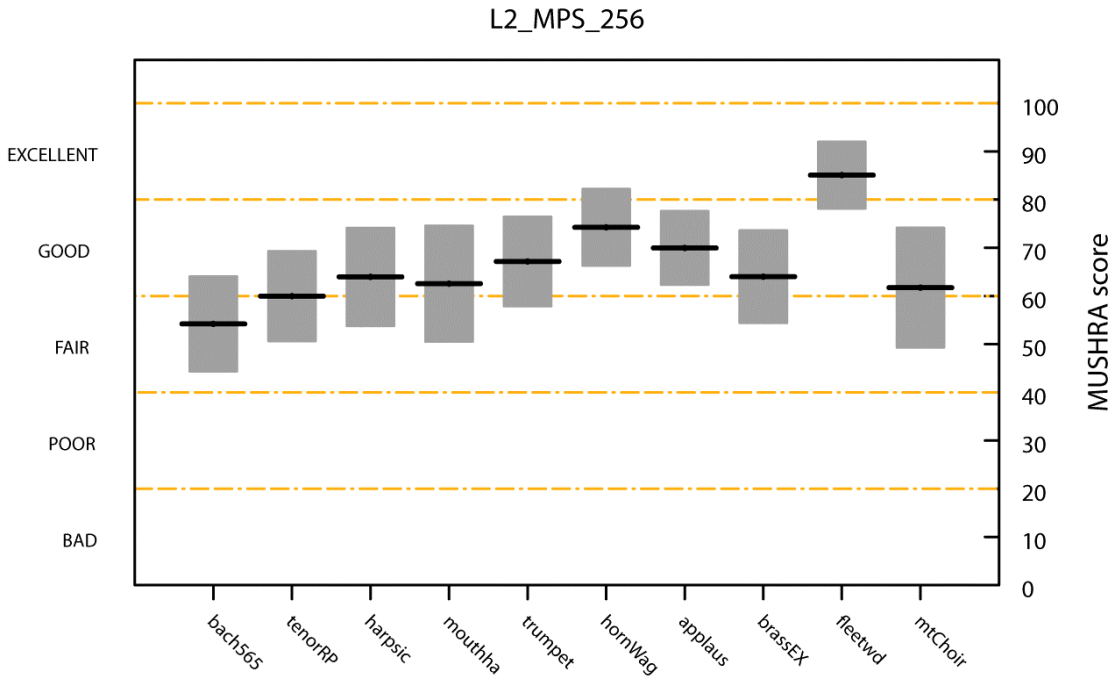

DD+_new_256
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir


DD+_old_200
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir


DD+_old_256
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir

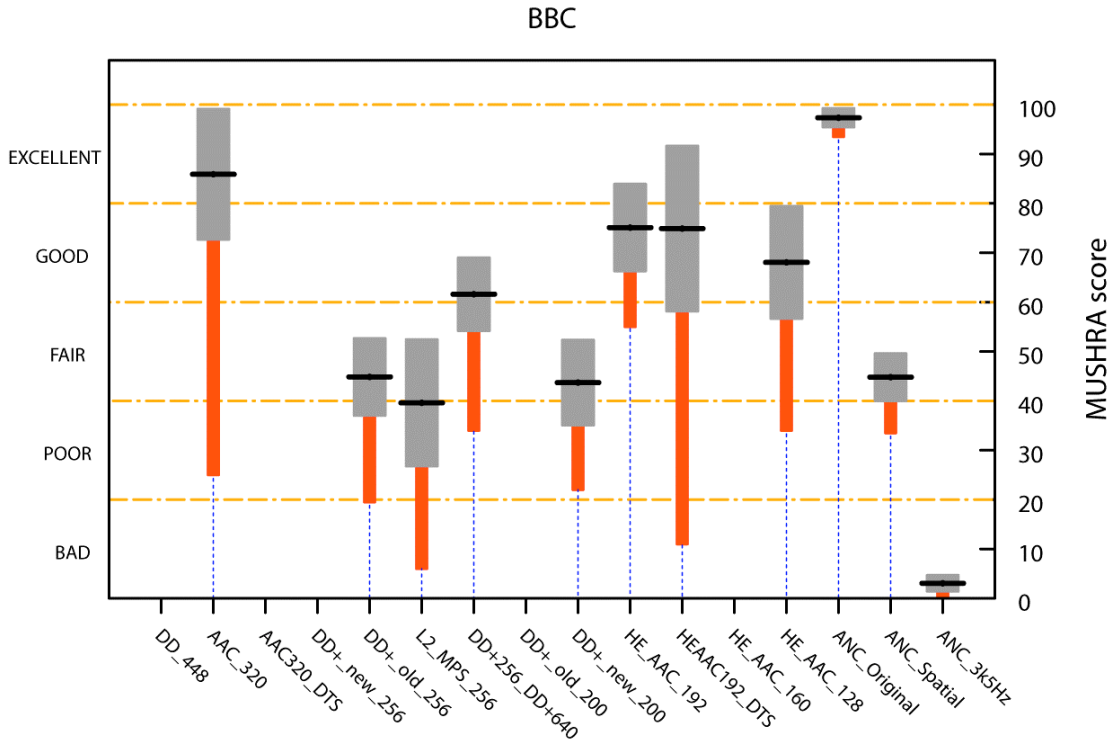

HE_AAC_128
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir
HE_AAC_160
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir

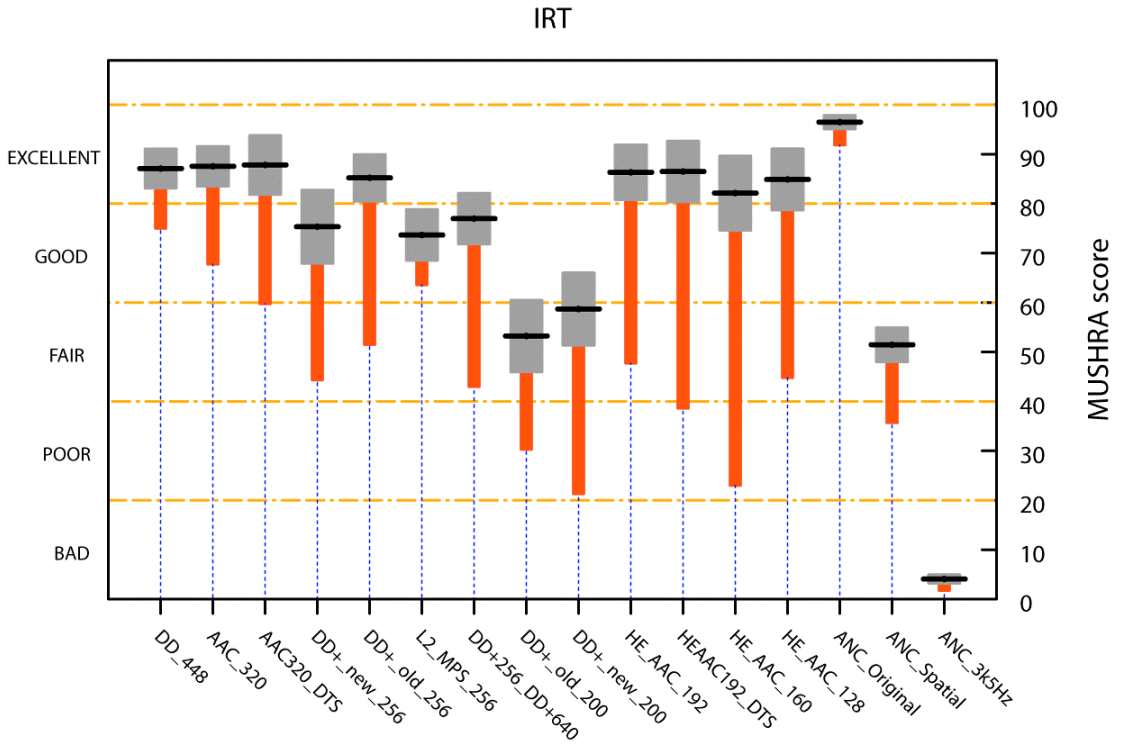

HE_AAC_192
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir
HEAAC192_DTS
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
bach565
tenorRP
harpsic
mouthha
trumpet
hornWag
applaus
brassEX
fleetwd
mtChoir

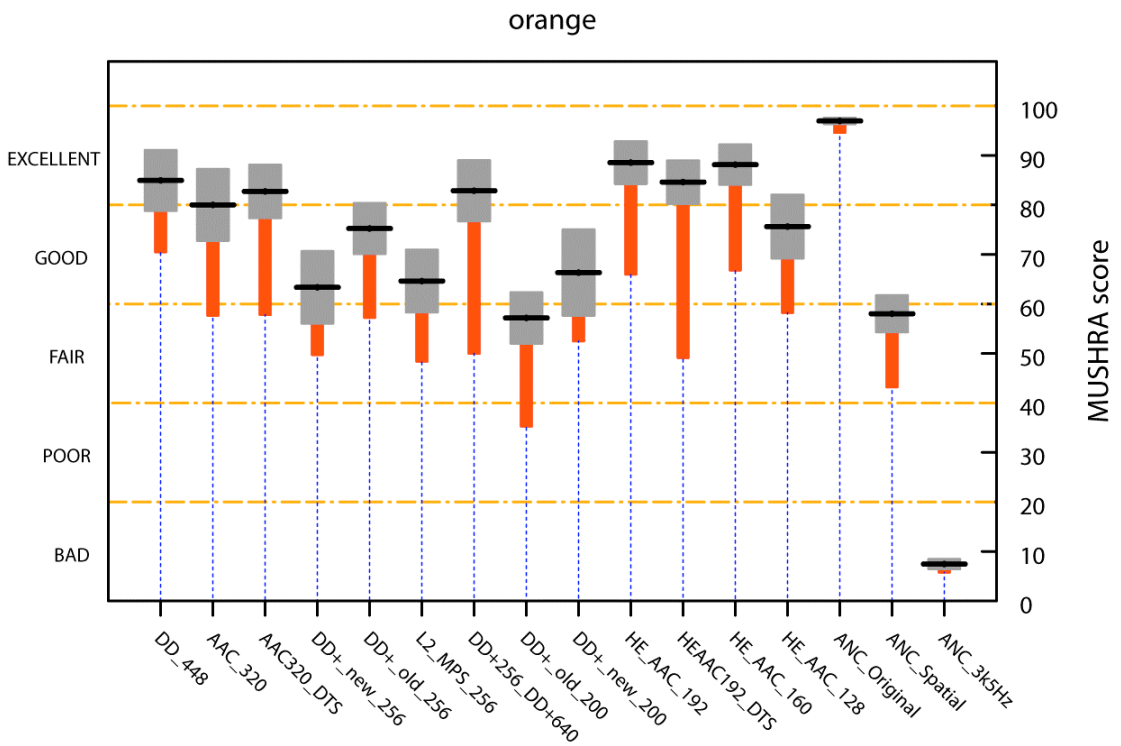


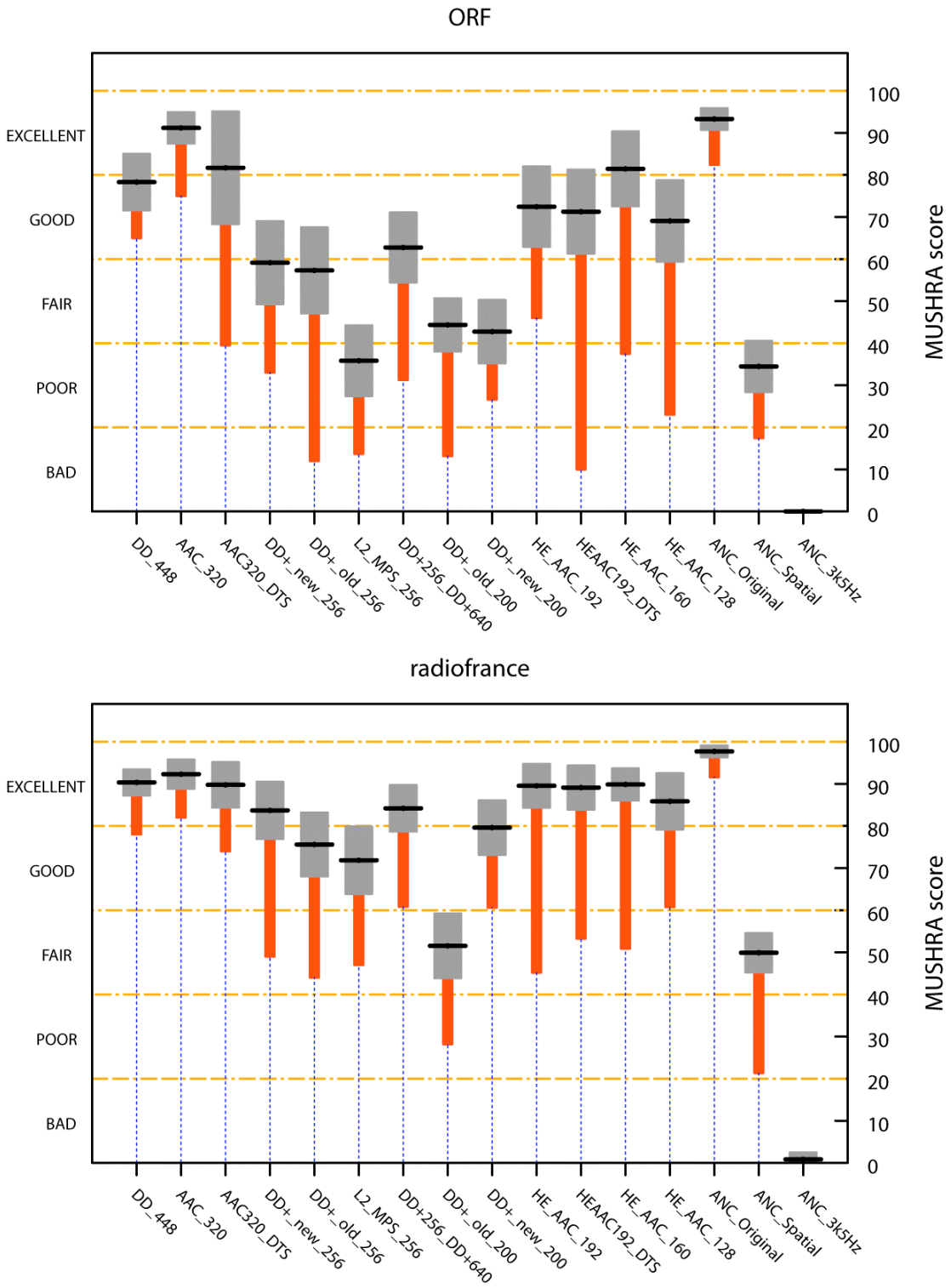
Average scores for each lab over all codecs - Phase 2

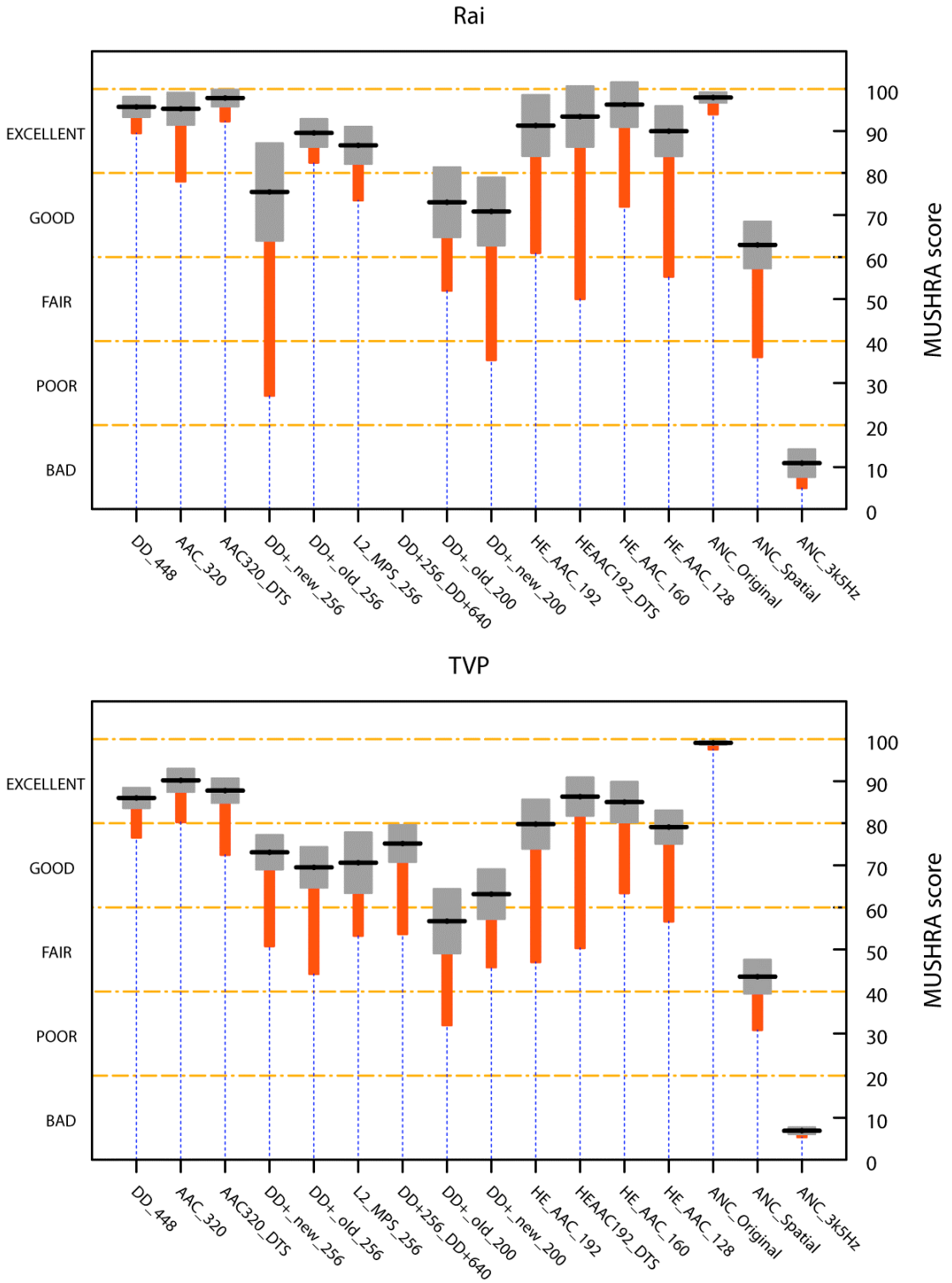
These graphs give the average and worse case scores for each codec on a lab by lab basis. While all the labs tested all the codecs, some listeners were rejected which resulted in some labs not having any scores for some codecs, hence the missing marks on some graphs.




IRT
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
DD_448
AAC_320
AAC320_DTS
DD+_new_256
DD+_old_256
L2_MPS_256
DD+256_DD+640
DD+_old_200
DD+_new_200
HE_AAC_192
HEAAC192_DTS
HE_AAC_160
HE_AAC_128
ANC_Original
ANC_Spatial
ANC_3k5Hz


orange
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
DD_448
AAC_320
AAC320_DTS
DD+_new_256
DD+_old_256
L2_MPS_256
DD+256_DD+640
DD+_old_200
DD+_new_200
HE_AAC_192
HEAAC192_DTS
HE_AAC_160
HE_AAC_128
ANC_Original
ANC_Spatial
ANC_3k5Hz

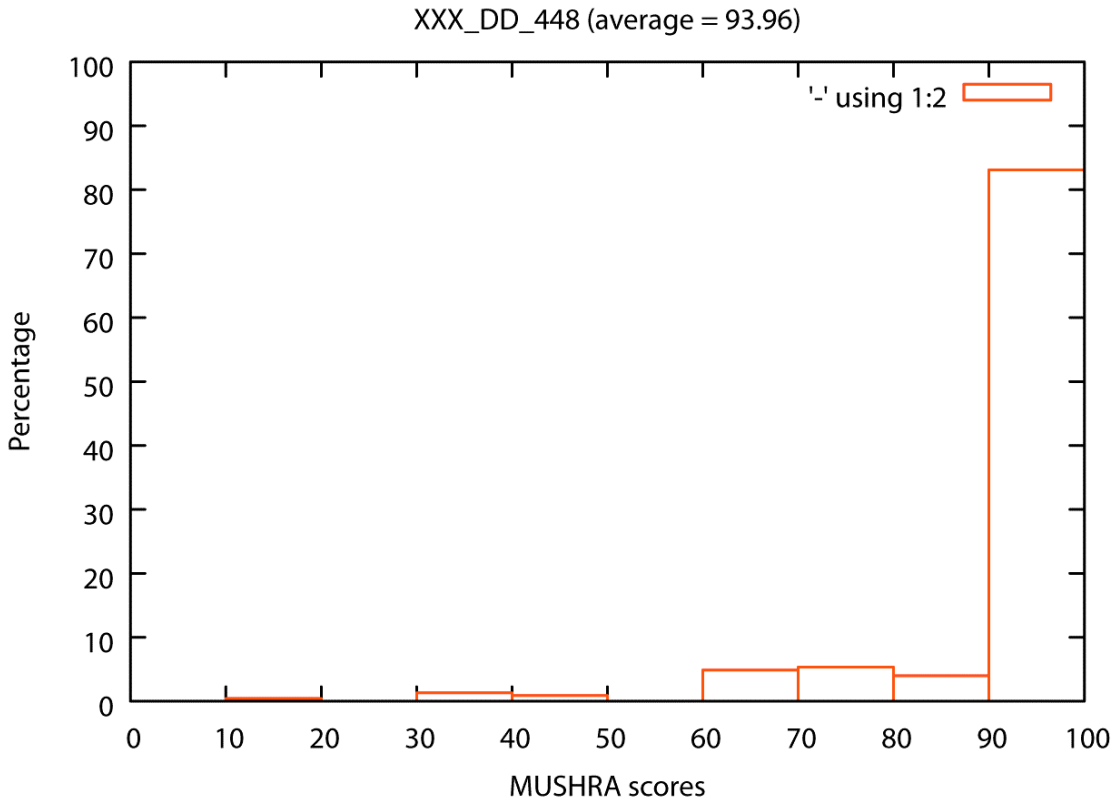

ORF
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
DD_448
AAC_320
AAC320_DTS
DD+_new_256
DD+_old_256
L2_MPS_256
DD+256_DD+640
DD+_old_200
DD+_new_200
HE_AAC_192
HEAAC192_DTS
HE_AAC_160
HE_AAC_128
ANC_Original
ANC_Spatial
ANC_3k5Hz
radiofrance
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
DD_448
AAC_320
AAC320_DTS
DD+_new_256
DD+_old_256
L2_MPS_256
DD+256_DD+640
DD+_old_200
DD+_new_200
HE_AAC_192
HEAAC192_DTS
HE_AAC_160
HE_AAC_128
ANC_Original
ANC_Spatial
ANC_3k5Hz

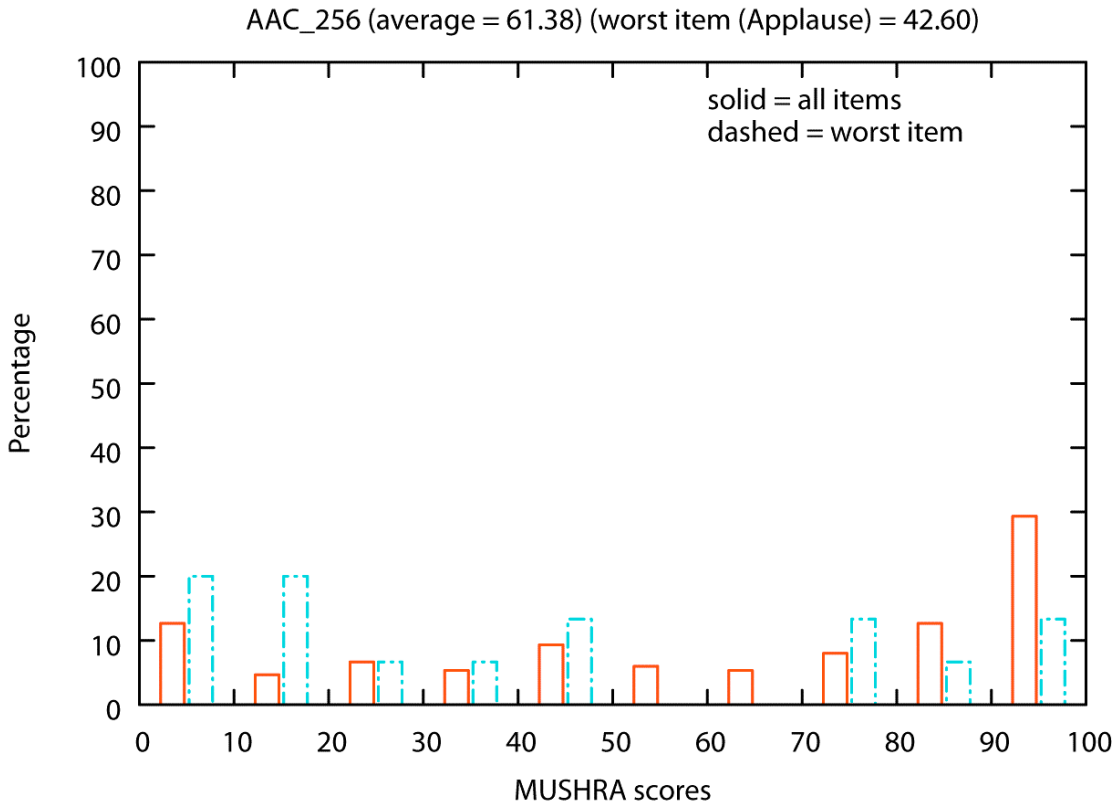

Rai
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
DD_448
AAC_320
AAC320_DTS
DD+_new_256
DD+_old_256
L2_MPS_256
DD+256_DD+640
DD+_old_200
DD+_new_200
HE_AAC_192
HEAAC192_DTS
HE_AAC_160
HE_AAC_128
ANC_Original
ANC_Spatial
ANC_3k5Hz
TVP
EXCELLENT
GOOD
FAIR
POOR
BAD
100
90
80
70
60
50
40
30
20
10
0
MUSHRA score
DD_448
AAC_320
AAC320_DTS
DD+_new_256
DD+_old_256
L2_MPS_256
DD+256_DD+640
DD+_old_200
DD+_new_200
HE_AAC_192
HEAAC192_DTS
HE_AAC_160
HE_AAC_128
ANC_Original
ANC_Spatial
ANC_3k5Hz

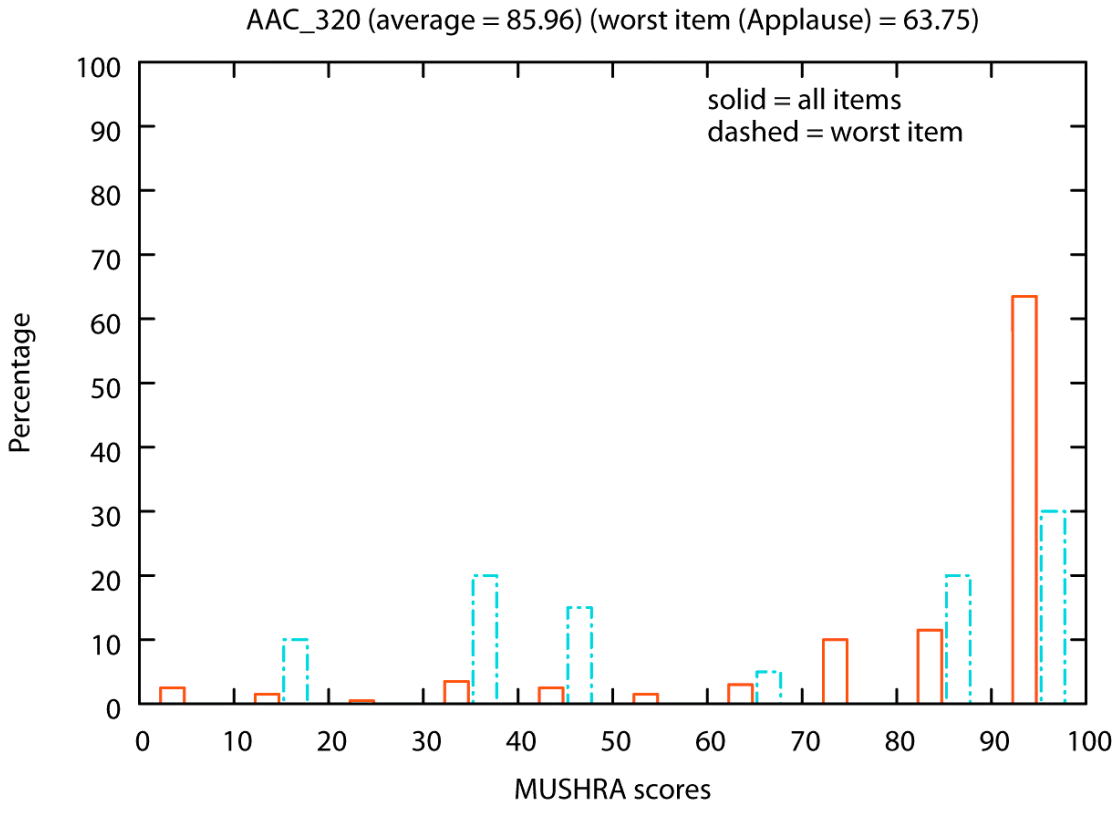
Appendix 5: Histograms

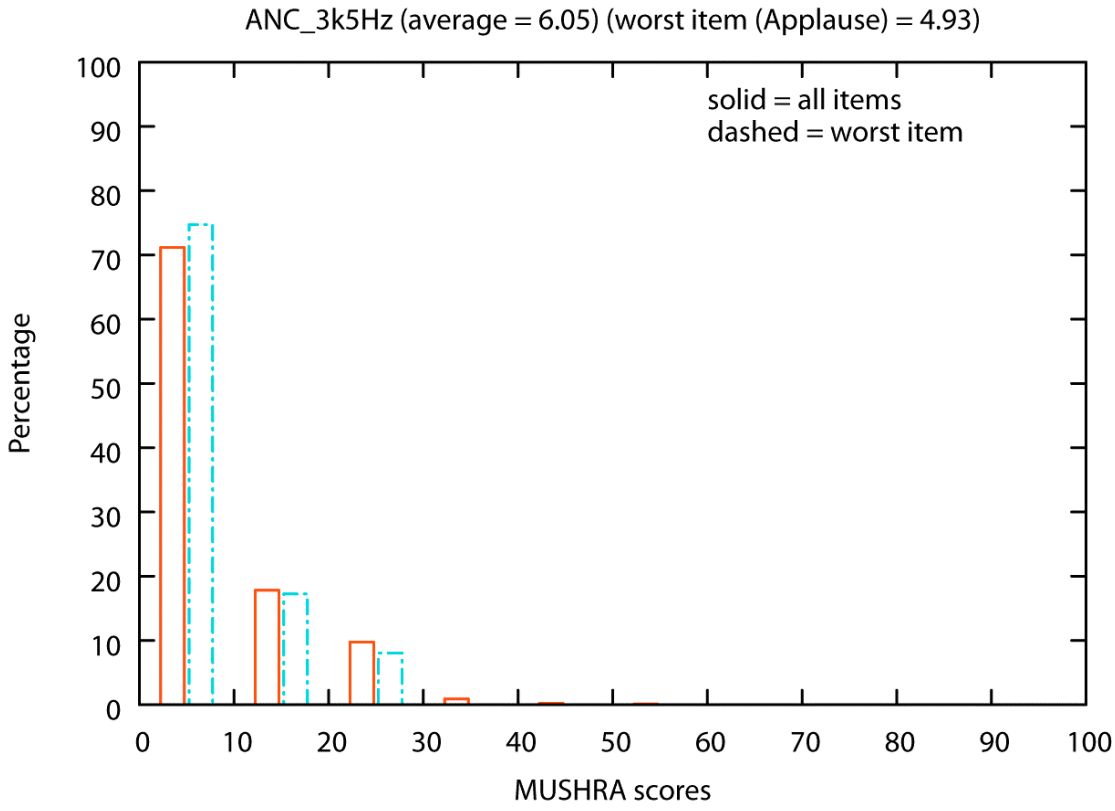
The distribution of scores for each codec are shown in these histograms. The red bars (on the left of the pair) indicate the scores over all the test items. The green bars (on the right of the pair) indicate the scores for the worst case item for that codec.

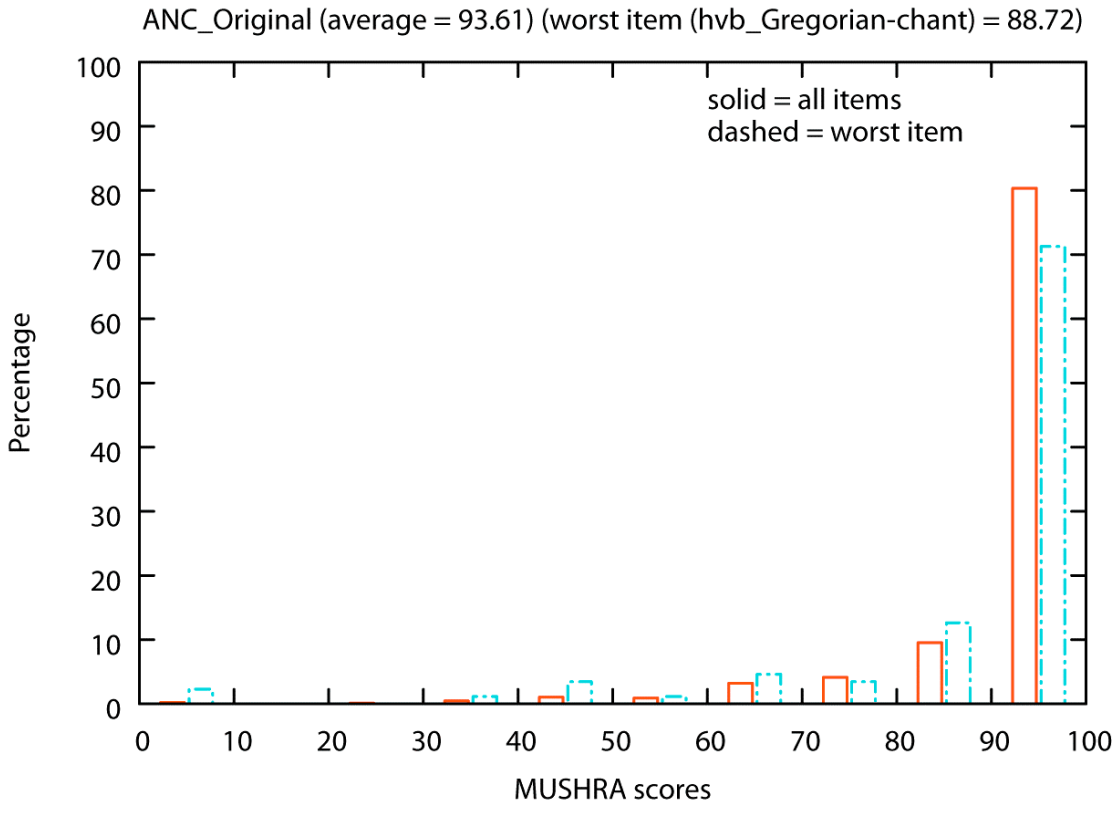
Phase 1

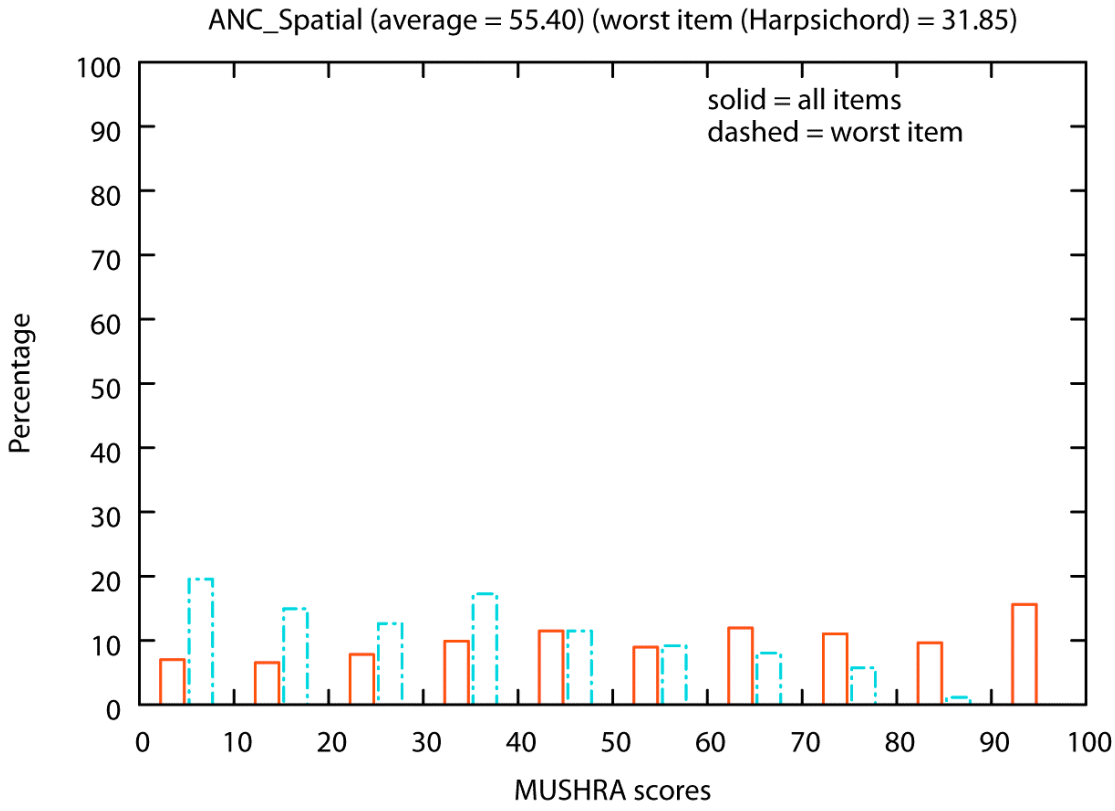


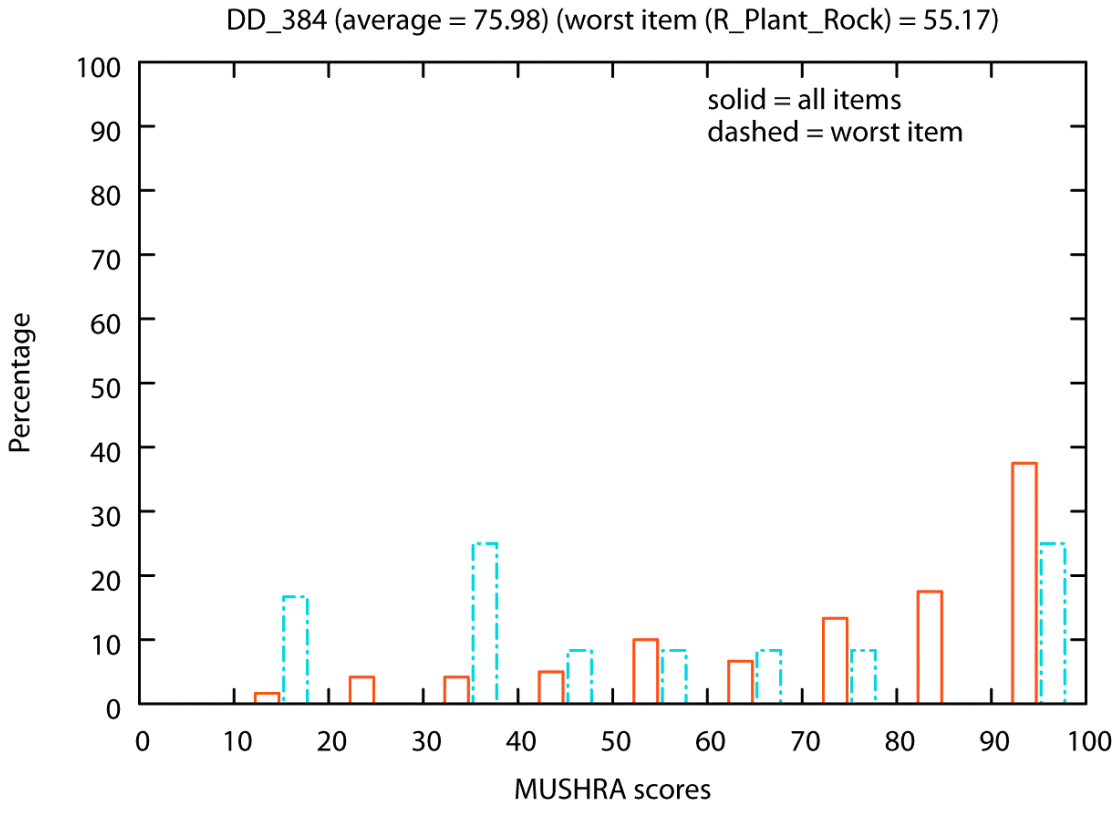

AAC_256 (average = 61.38) (worst item (Applause) = 42.60)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

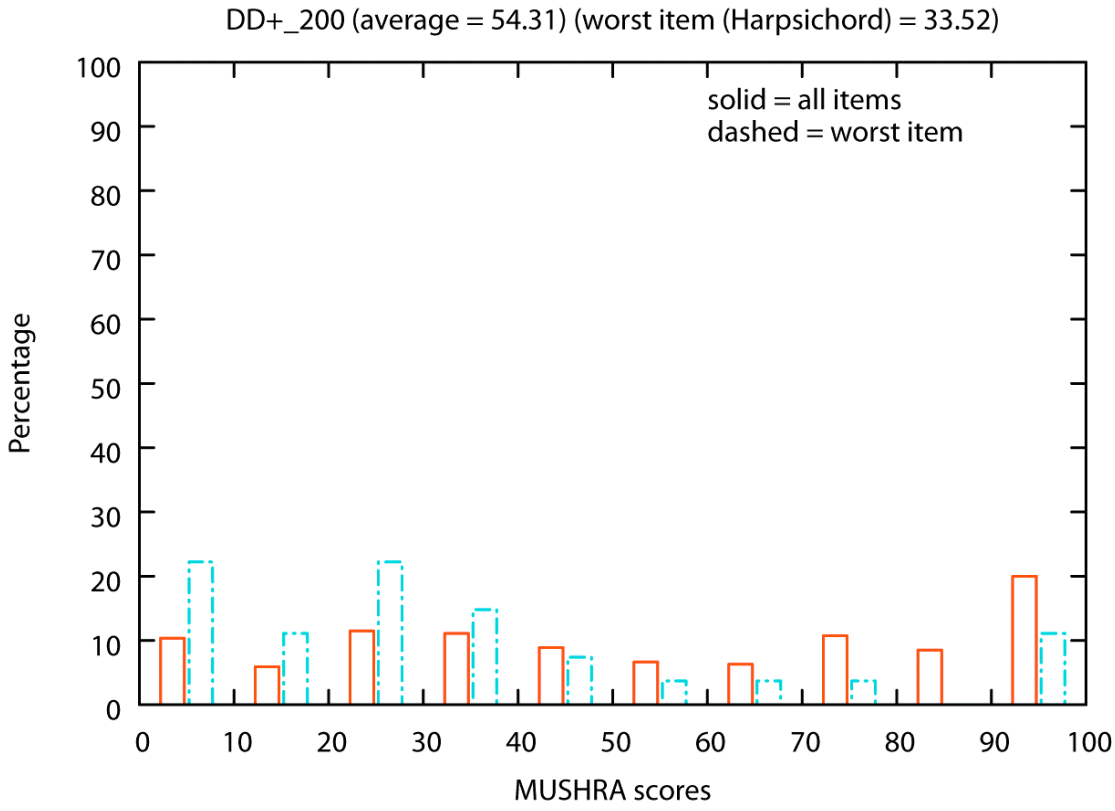

AAC_320 (average = 85.96) (worst item (Applause) = 63.75)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

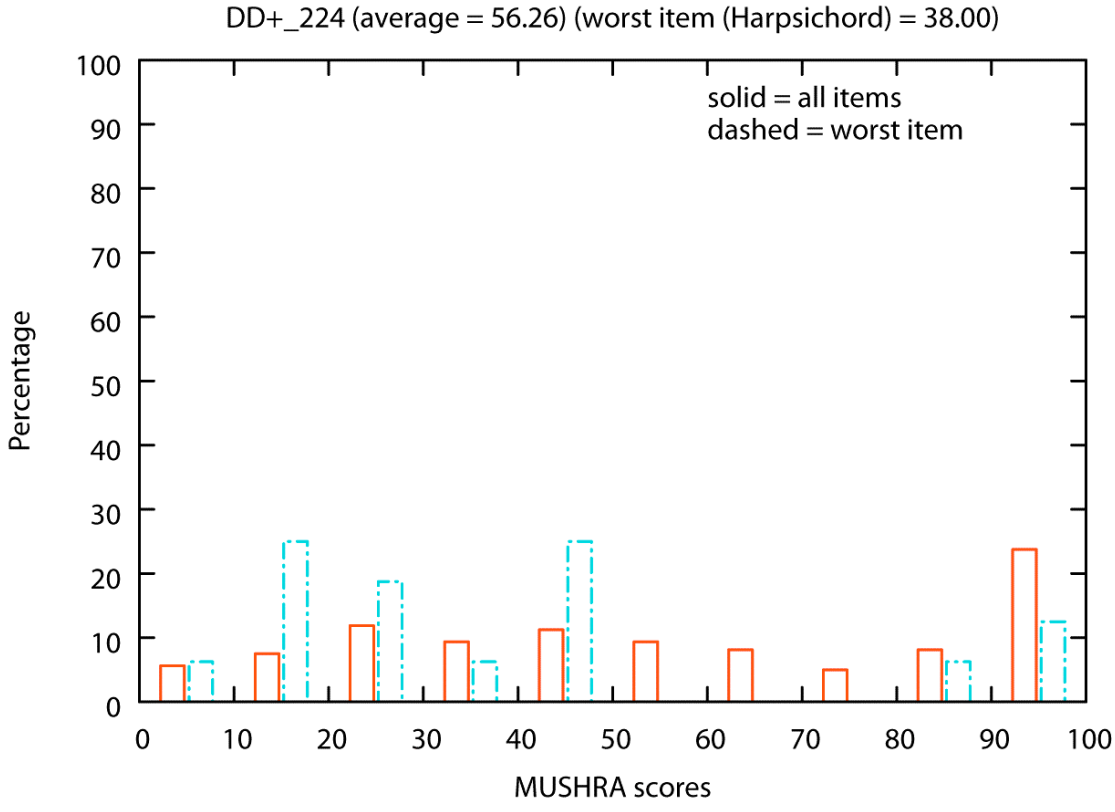

ANC_3k5Hz (average = 6.05) (worst item (Applause) = 4.93)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

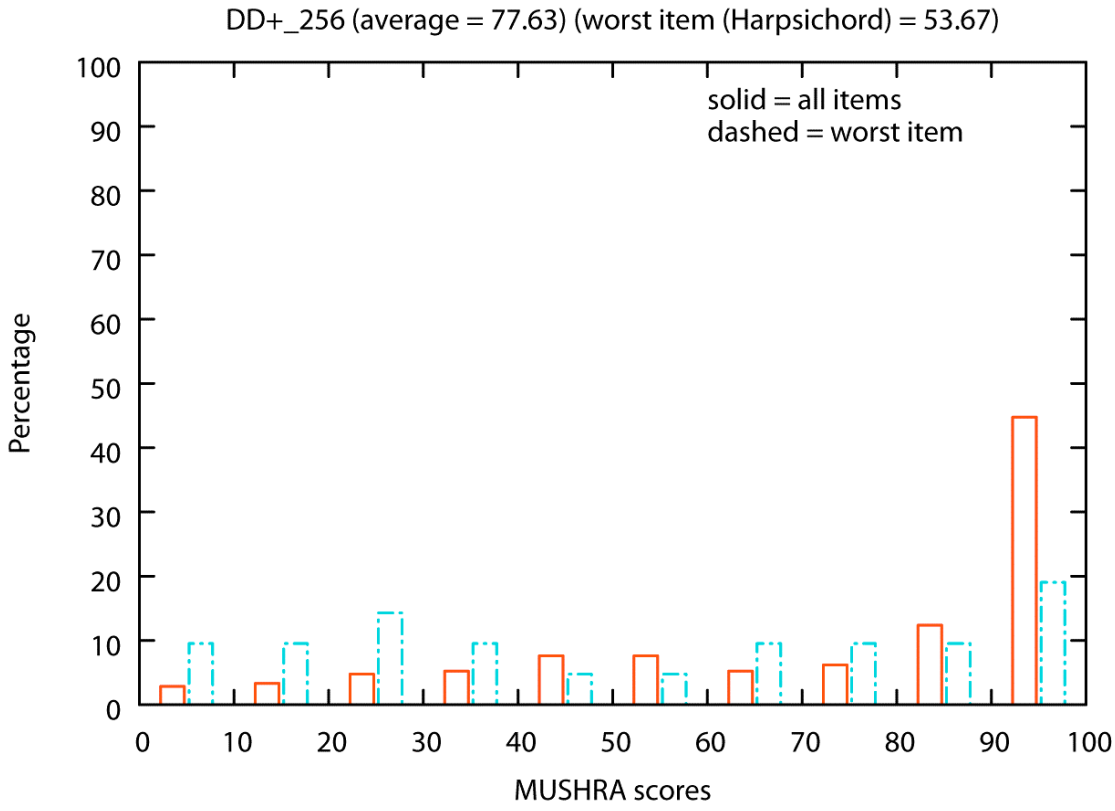

ANC_Original (average = 93.61) (worst item (hvb_Gregorian-chant) = 88.72)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

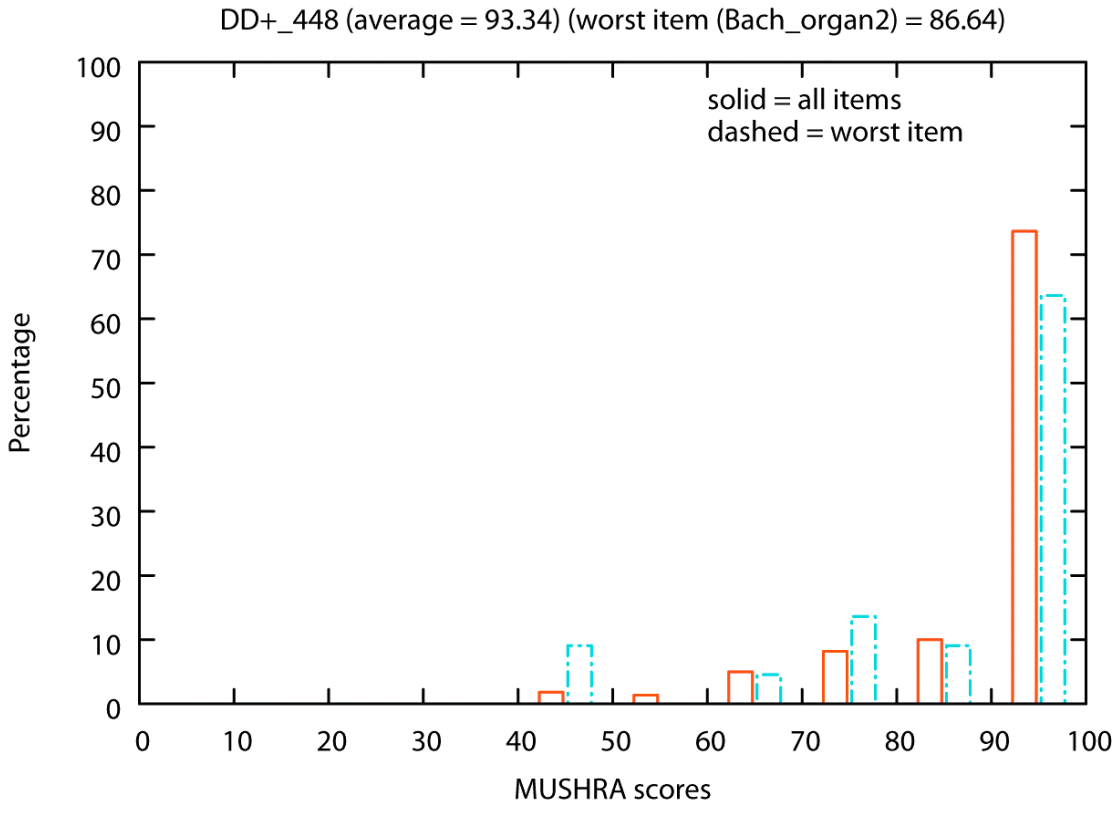

ANC_Spatial (average = 55.40) (worst item (Harpsichord) = 31.85)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

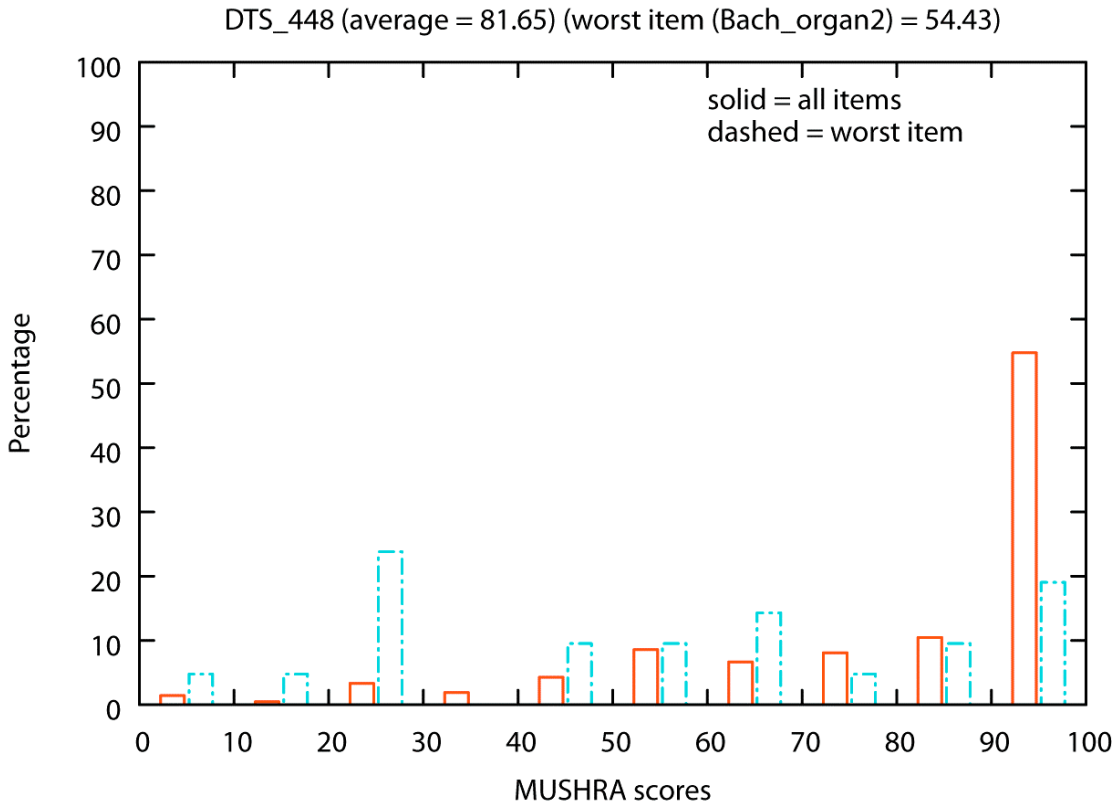

DD_384 (average = 75.98) (worst item (R_Plant_Rock) = 55.17)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

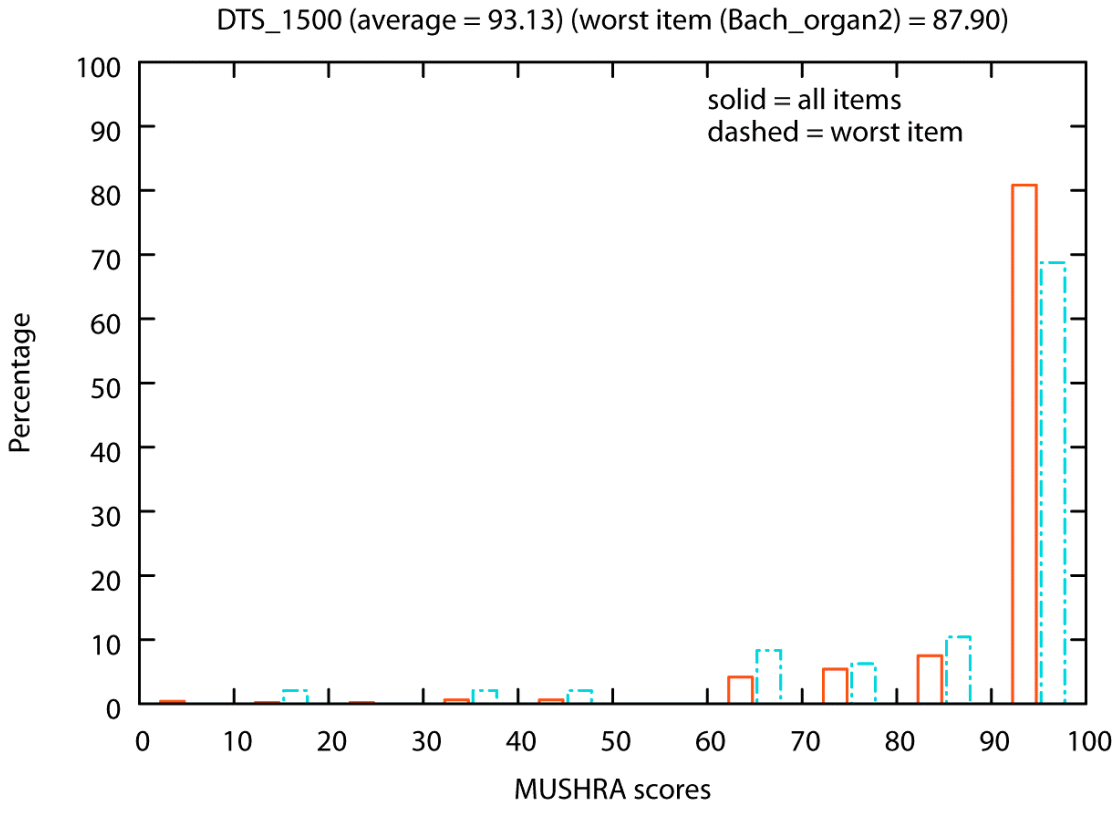

DD+_200 (average = 54.31) (worst item (Harpsichord) = 33.52)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

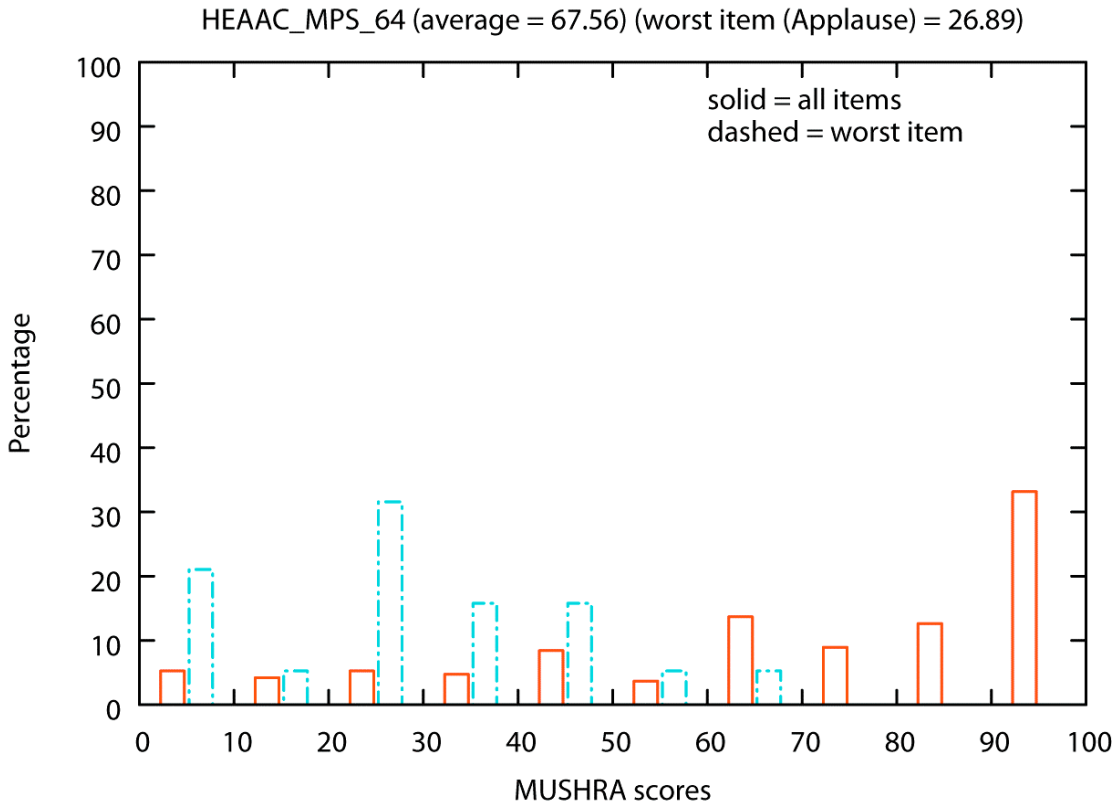

DD+_224 (average = 56.26) (worst item (Harpsichord) = 38.00)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

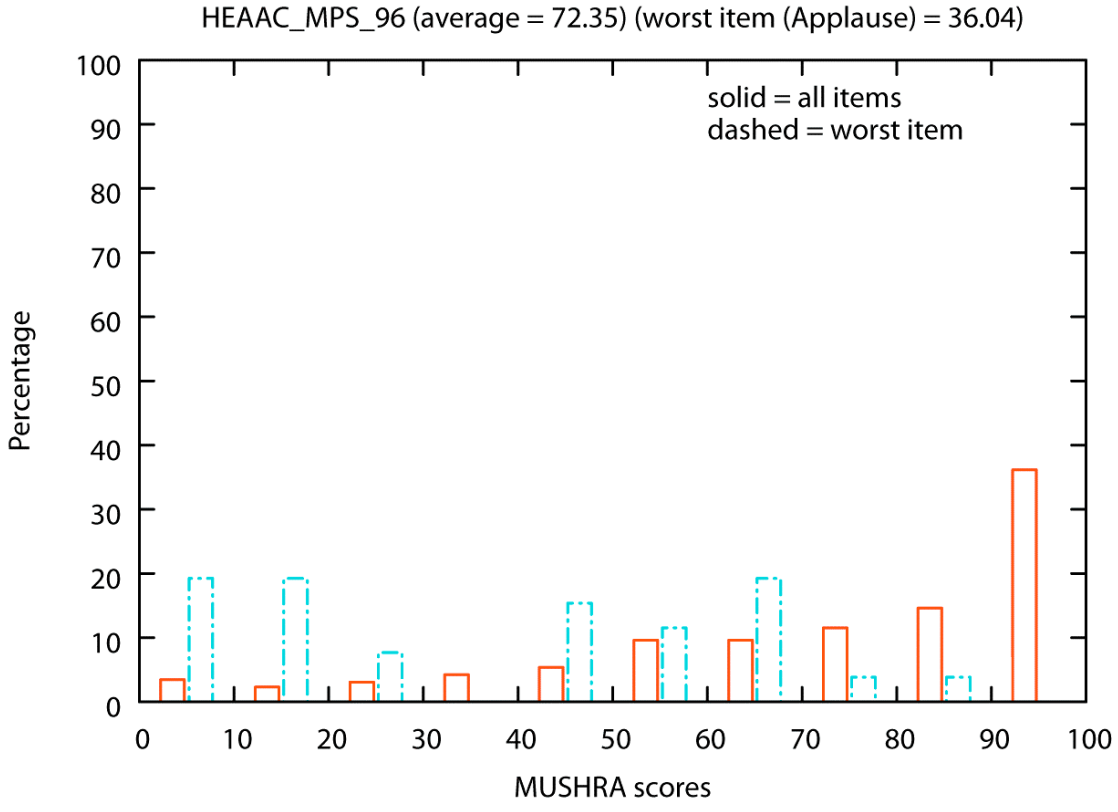

DD+_256 (average = 77.63) (worst item (Harpsichord) = 53.67)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

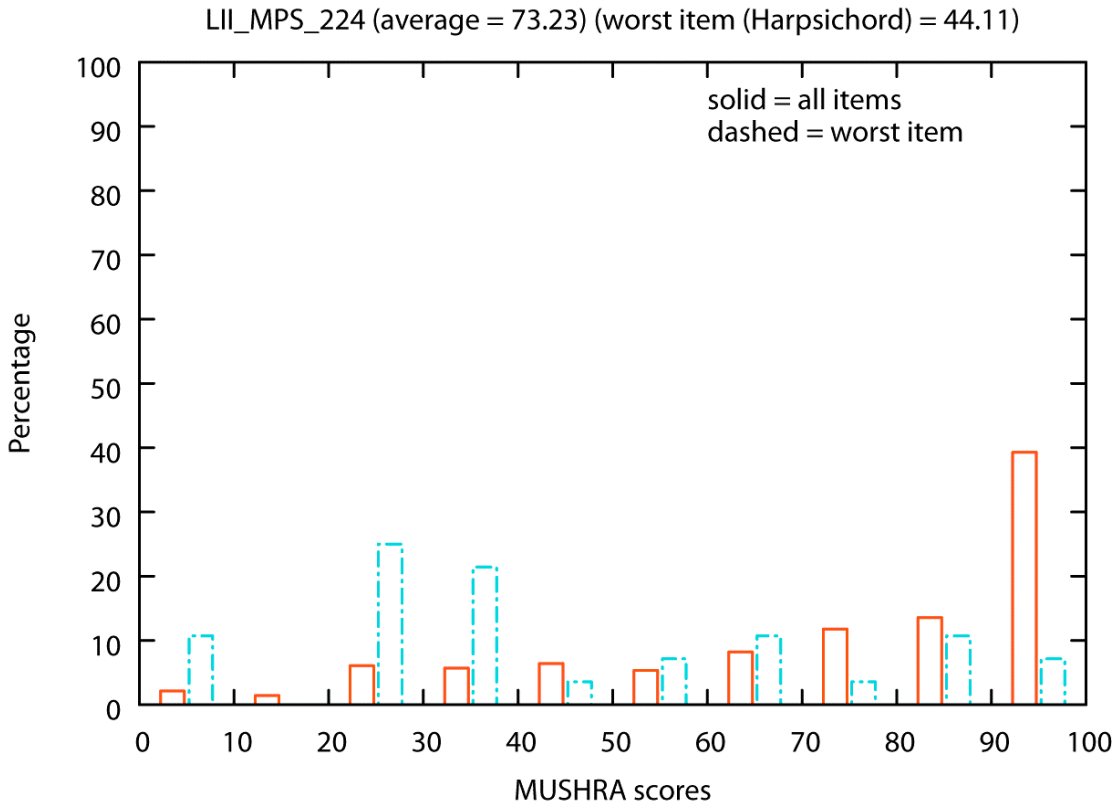

DD+_448 (average = 93.34) (worst item (Bach_organ2) = 86.64)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

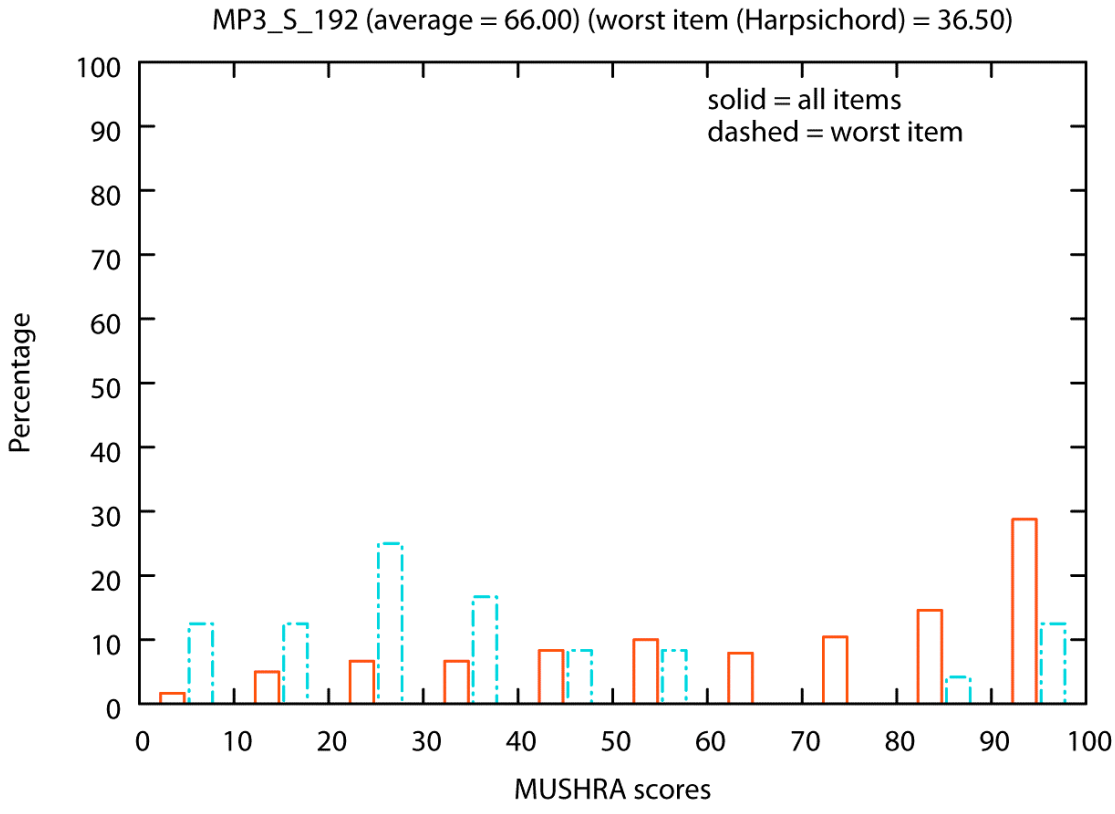

DTS_448 (average = 81.65) (worst item (Bach_organ2) = 54.43)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

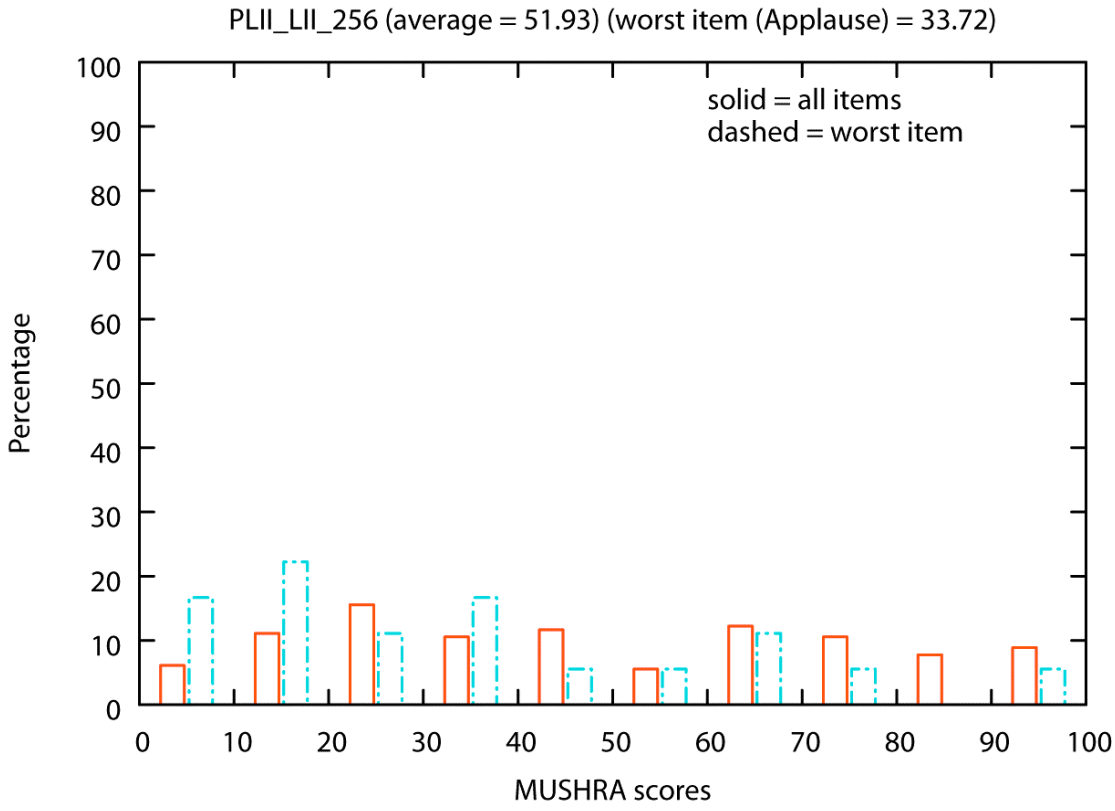

DTS_1500 (average = 93.13) (worst item (Bach_organ2) = 87.90)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

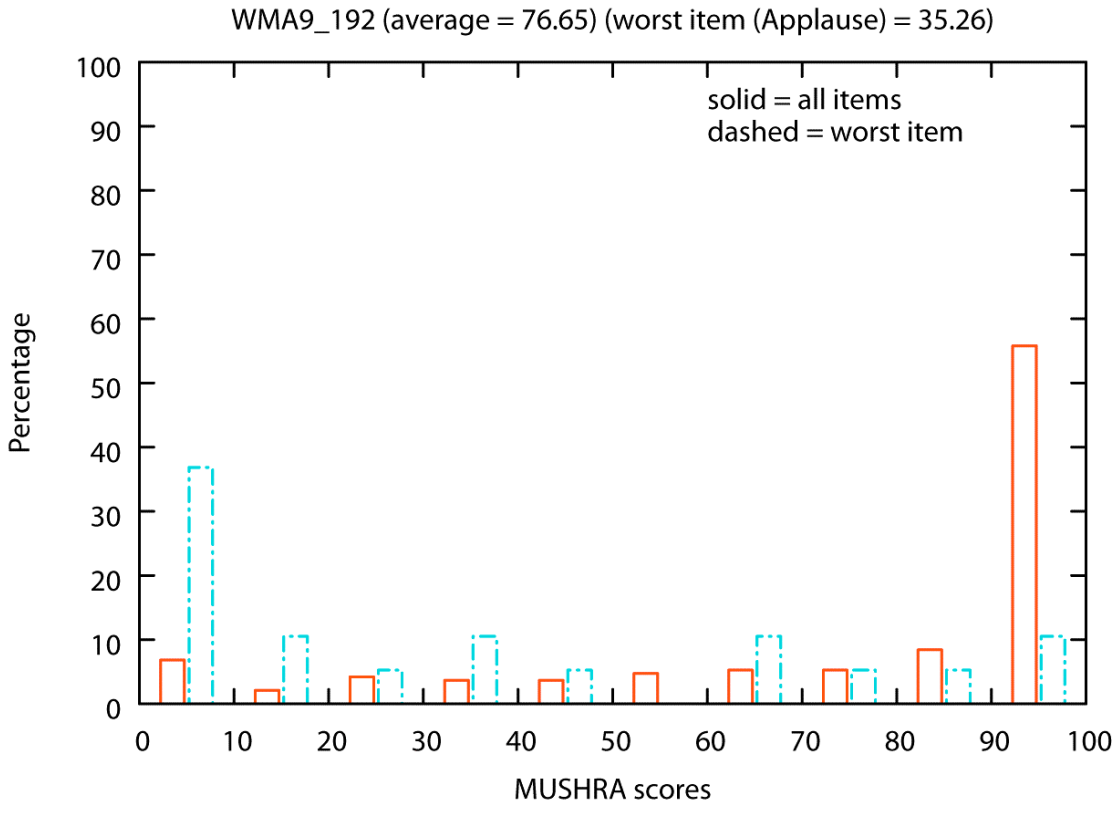

HEAAC_MPS_64 (average = 67.56) (worst item (Applause) = 26.89)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

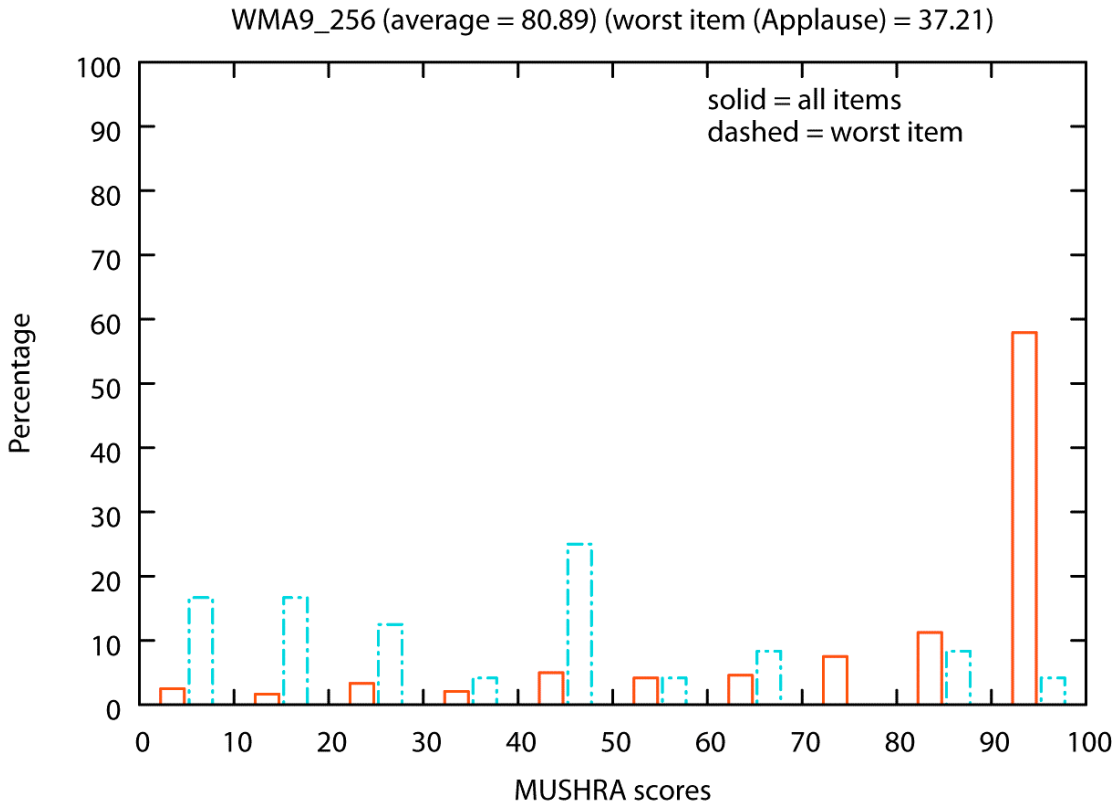

HEAAC_MPS_96 (average = 72.35) (worst item (Applause) = 36.04)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

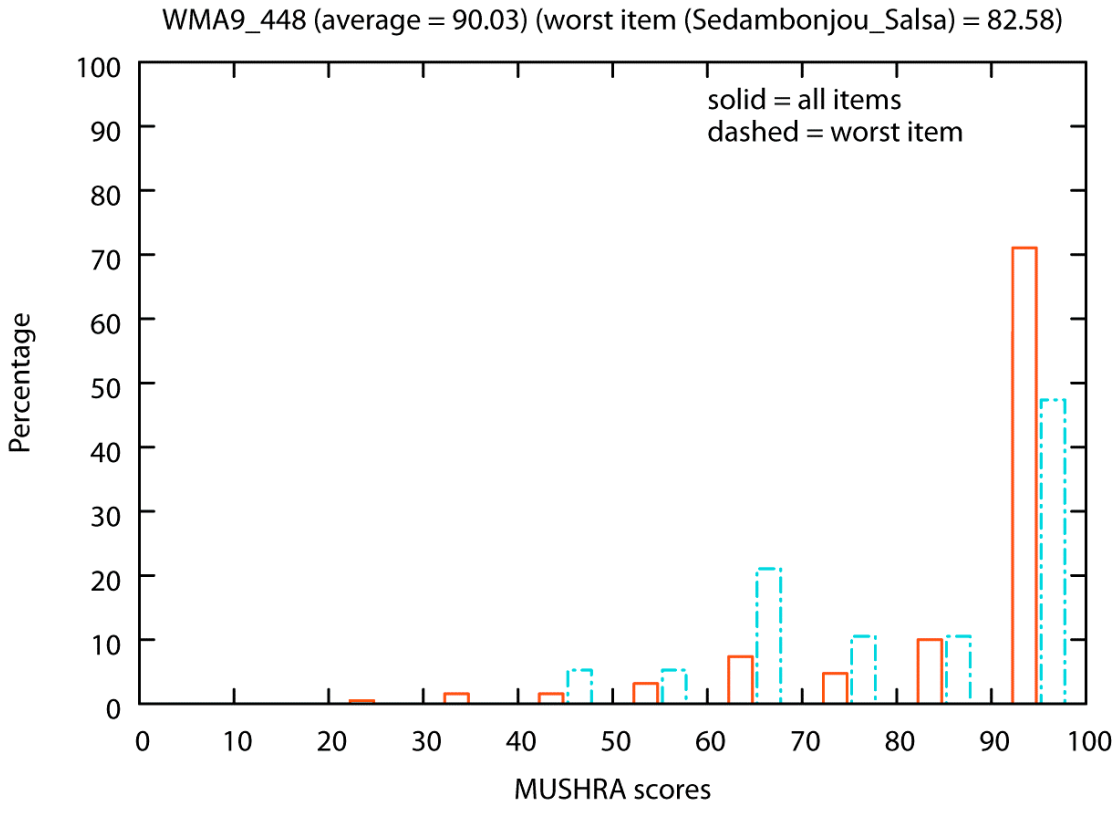

LII_MPS_224 (average = 73.23) (worst item (Harpsichord) = 44.11)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

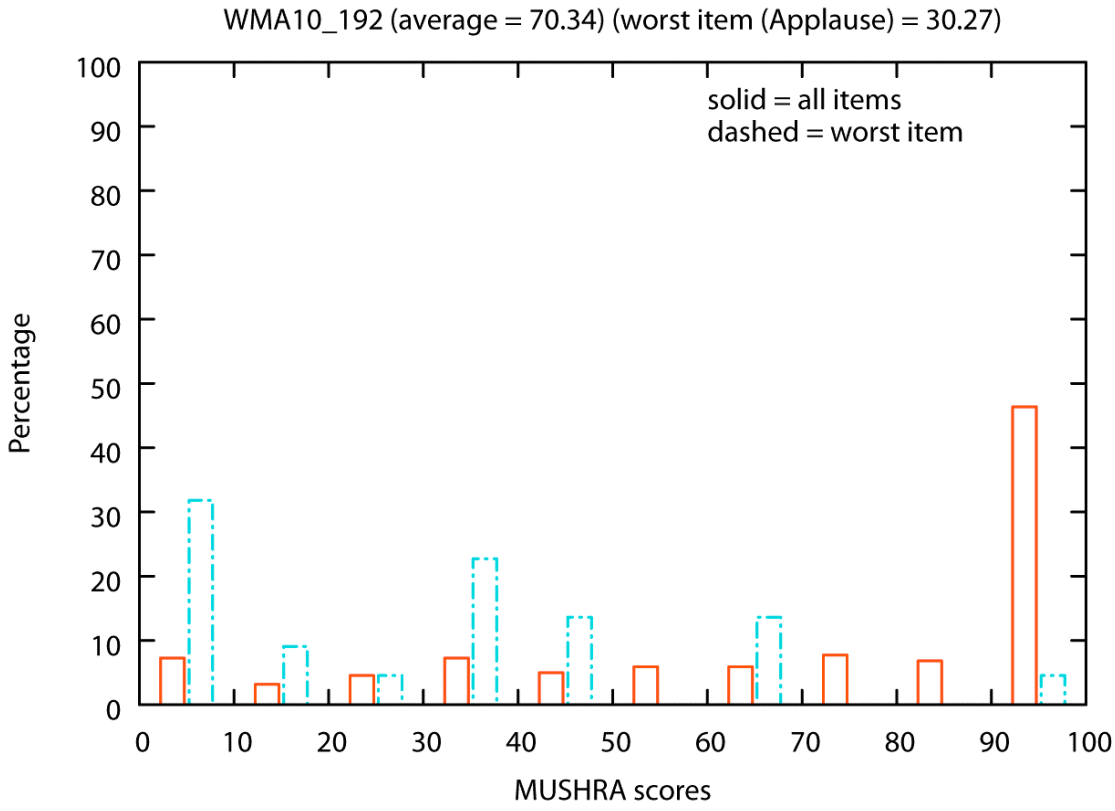

MP3_S_192 (average = 66.00) (worst item (Harpsichord) = 36.50)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

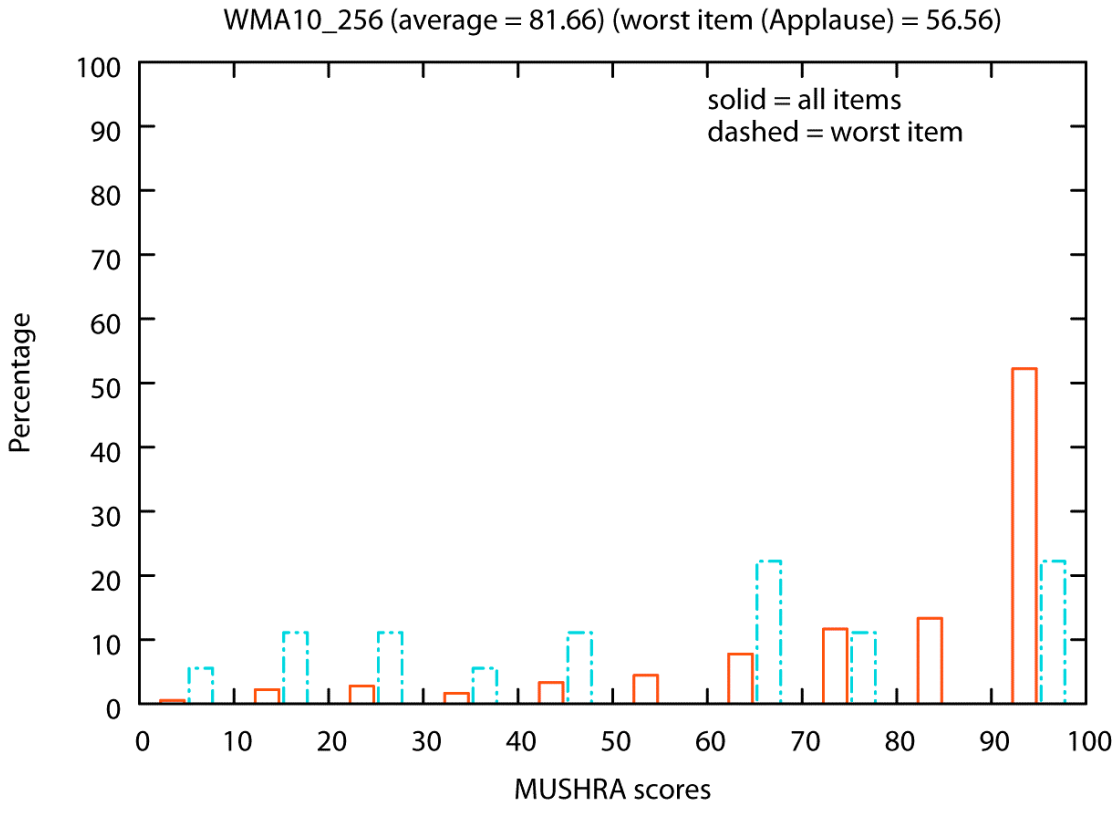

PLII_LII_256 (average = 51.93) (worst item (Applause) = 33.72)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100


WMA9_192 (average = 76.65) (worst item (Applause) = 35.26)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

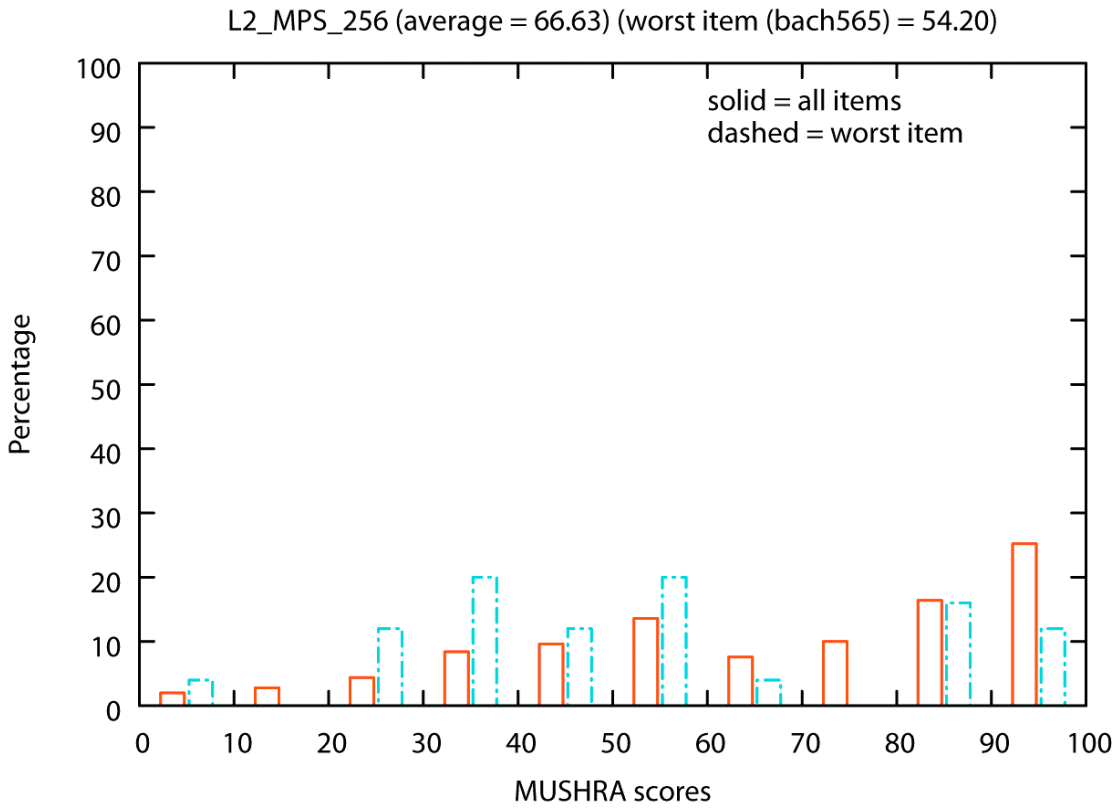

WMA9_256 (average = 80.89) (worst item (Applause) = 37.21)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

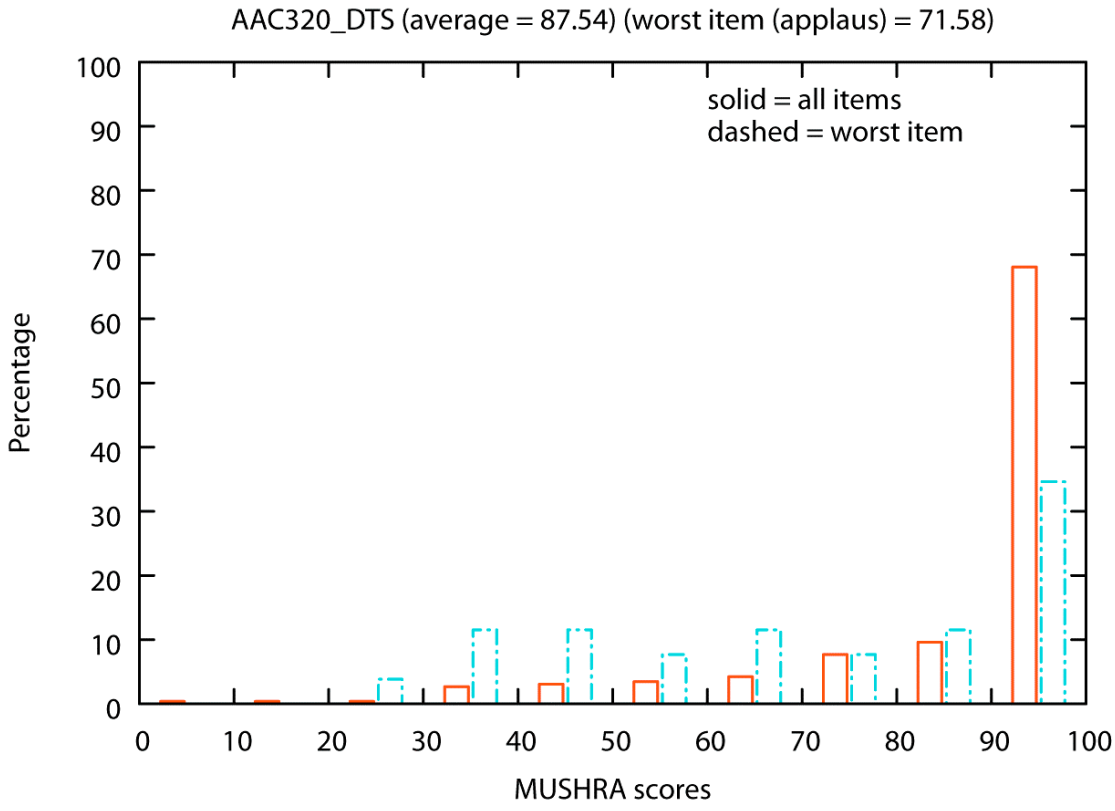

WMA9_448 (average = 90.03) (worst item (Sedambonjou_Salsa) = 82.58)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

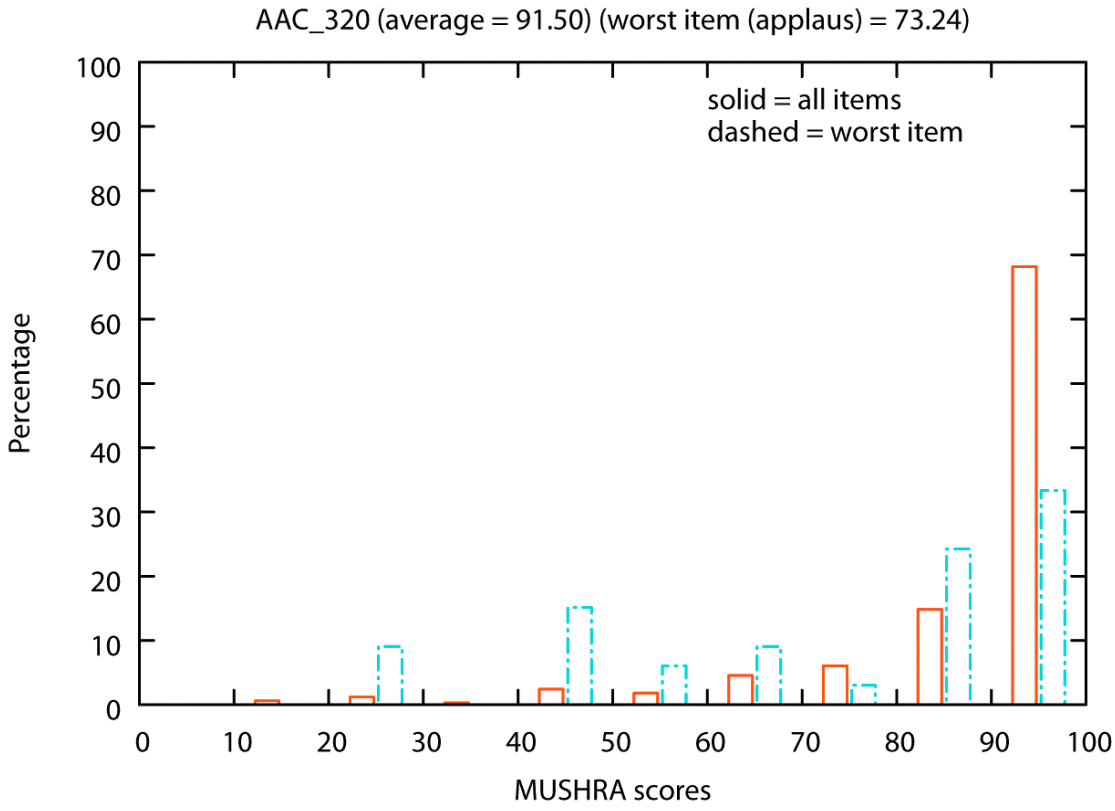

WMA10_192 (average = 70.34) (worst item (Applause) = 30.27)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

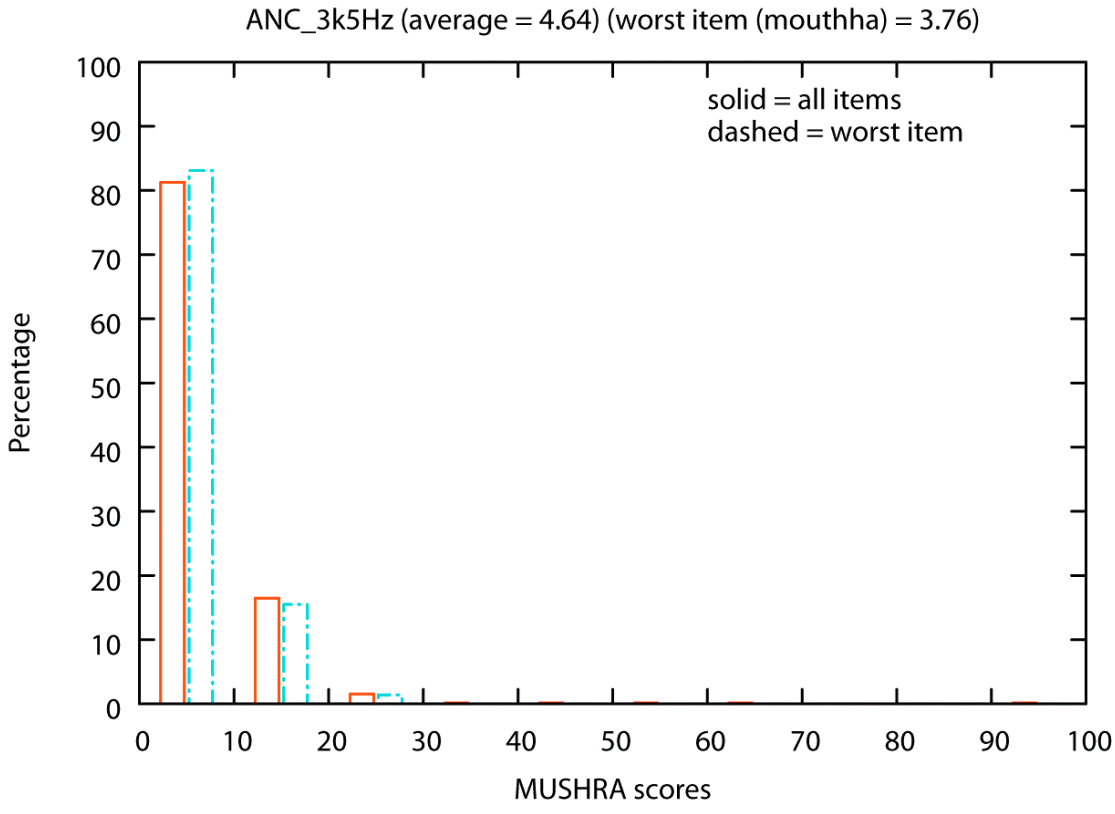

WMA10_256 (average = 81.66) (worst item (Applause) = 56.56)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

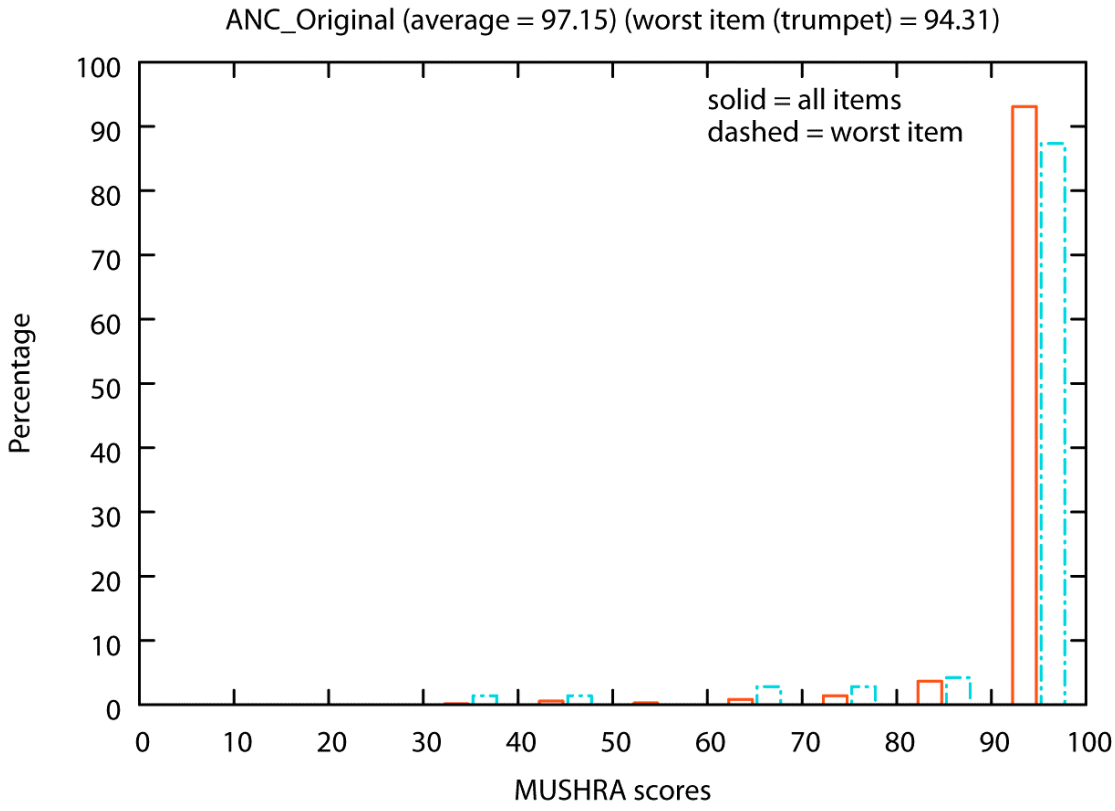
Phase 2

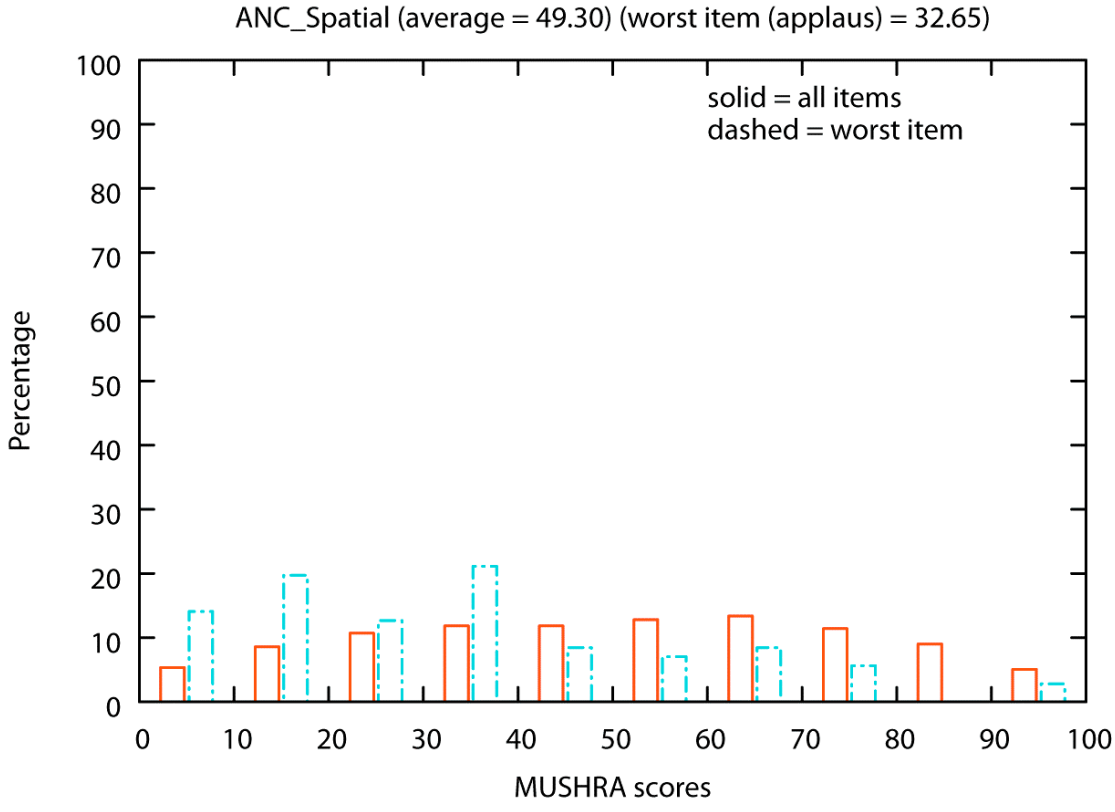

L2_MPS_256 (average = 66.63) (worst item (bach565) = 54.20)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

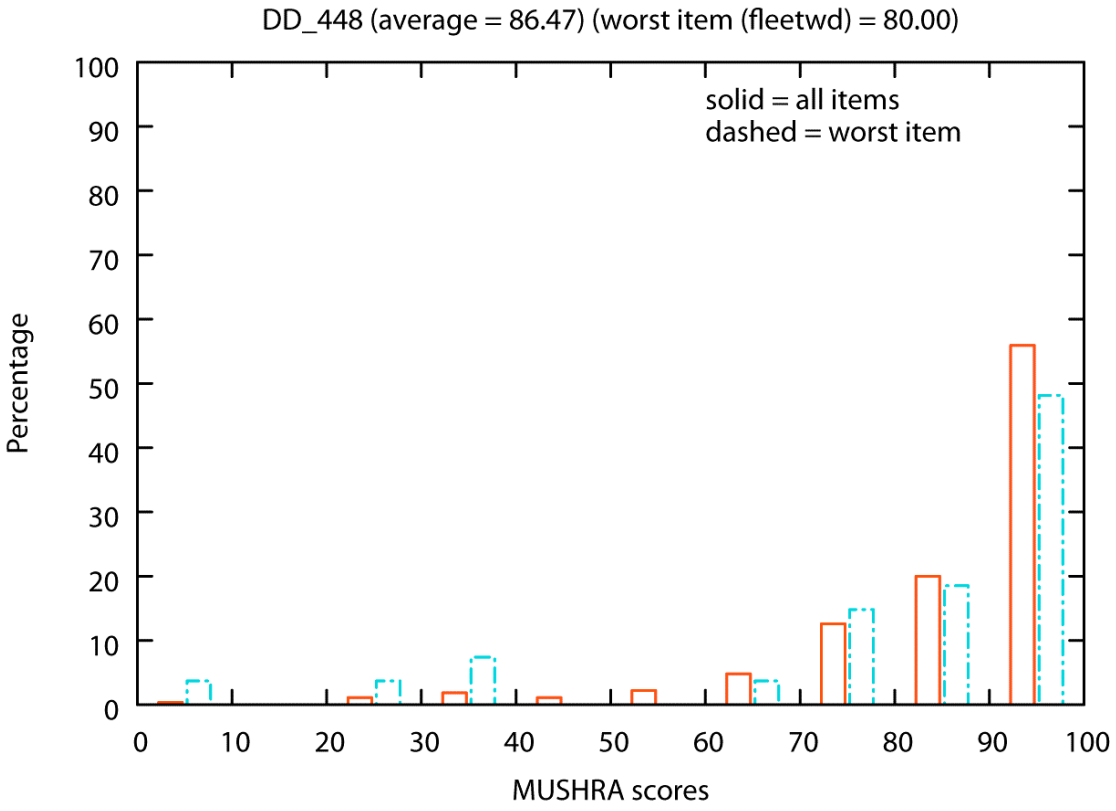

AAC320_DTS (average = 87.54) (worst item (applaus) = 71.58)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

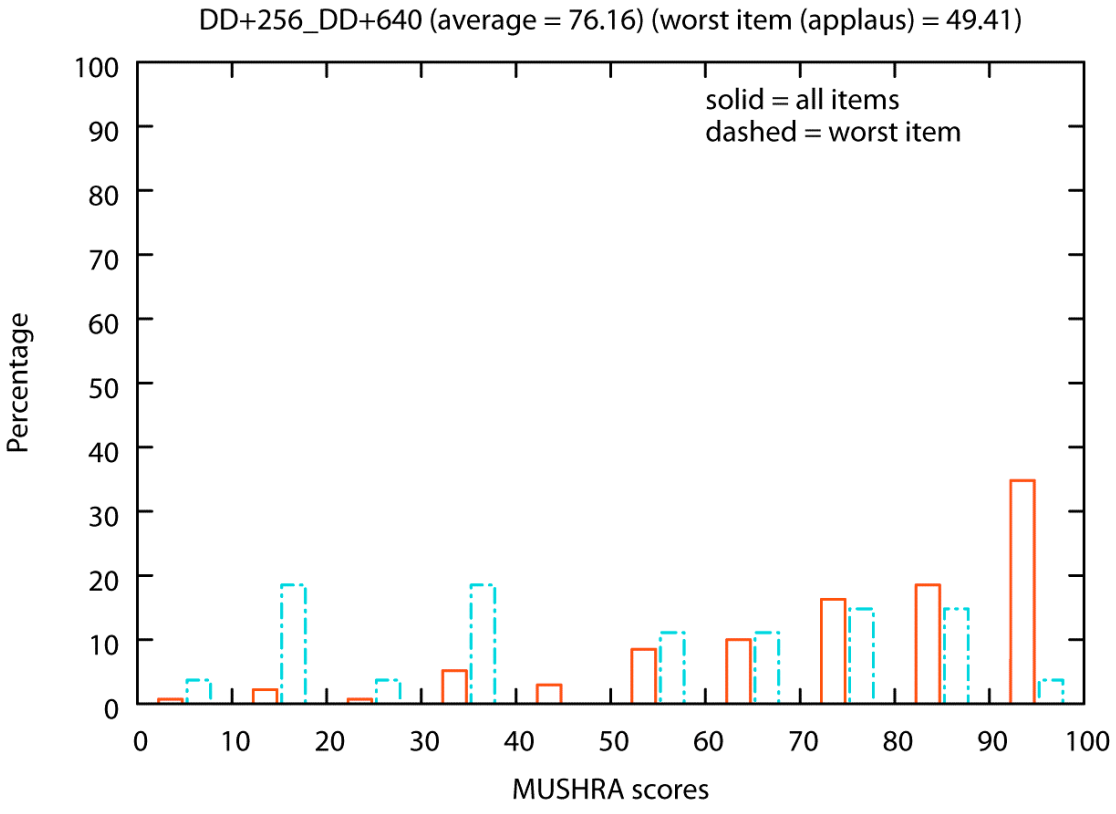

AAC_320 (average = 91.50) (worst item (applaus) = 73.24)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

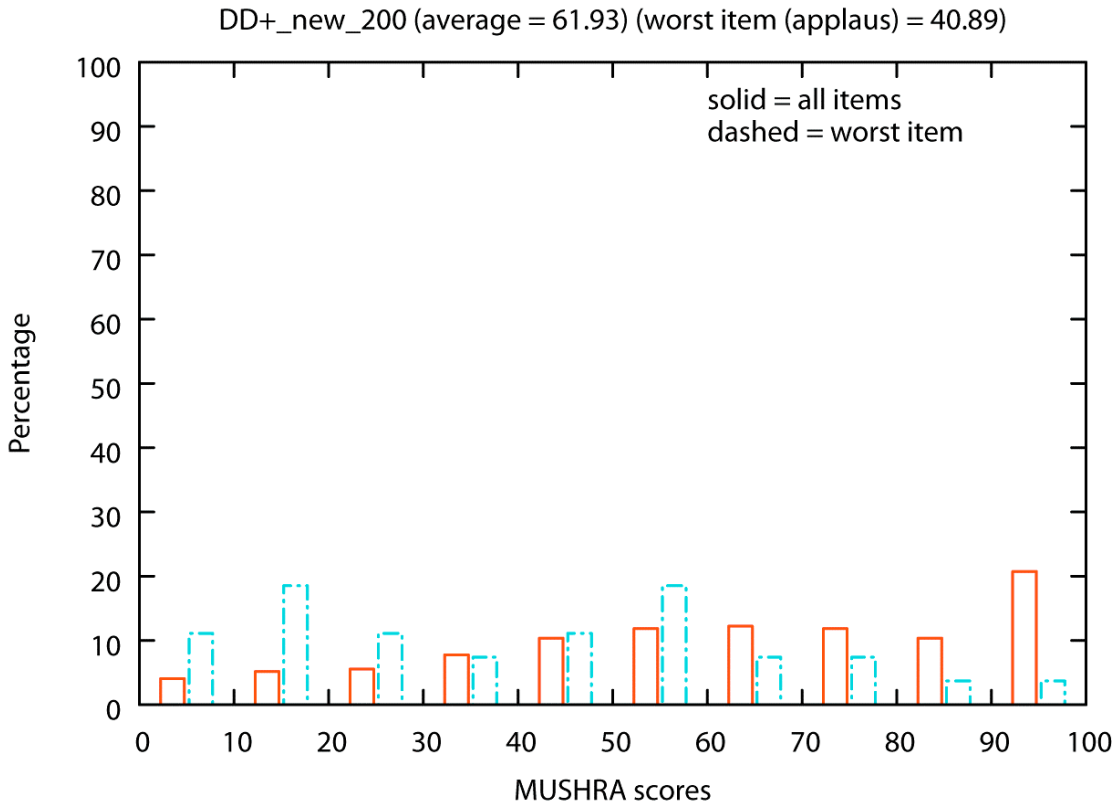

ANC_3k5Hz (average = 4.64) (worst item (mouthha) = 3.76)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

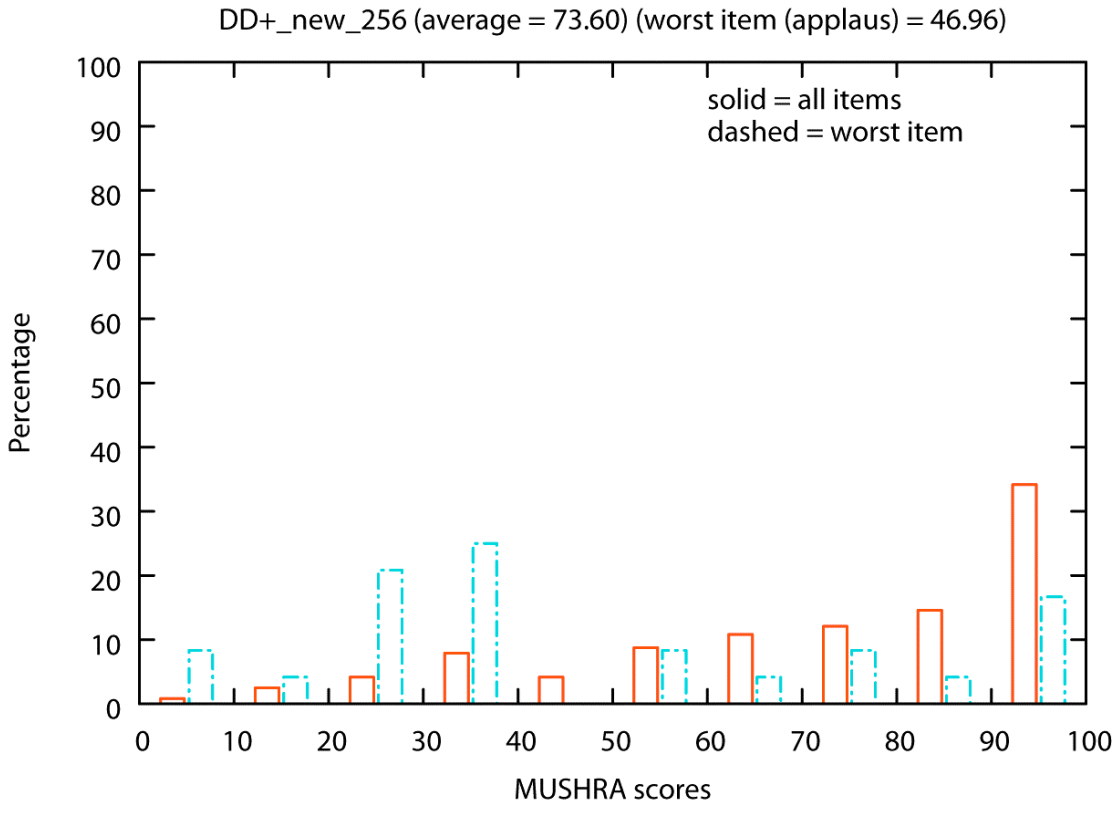

ANC_Original (average = 97.15) (worst item (trumpet) = 94.31)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

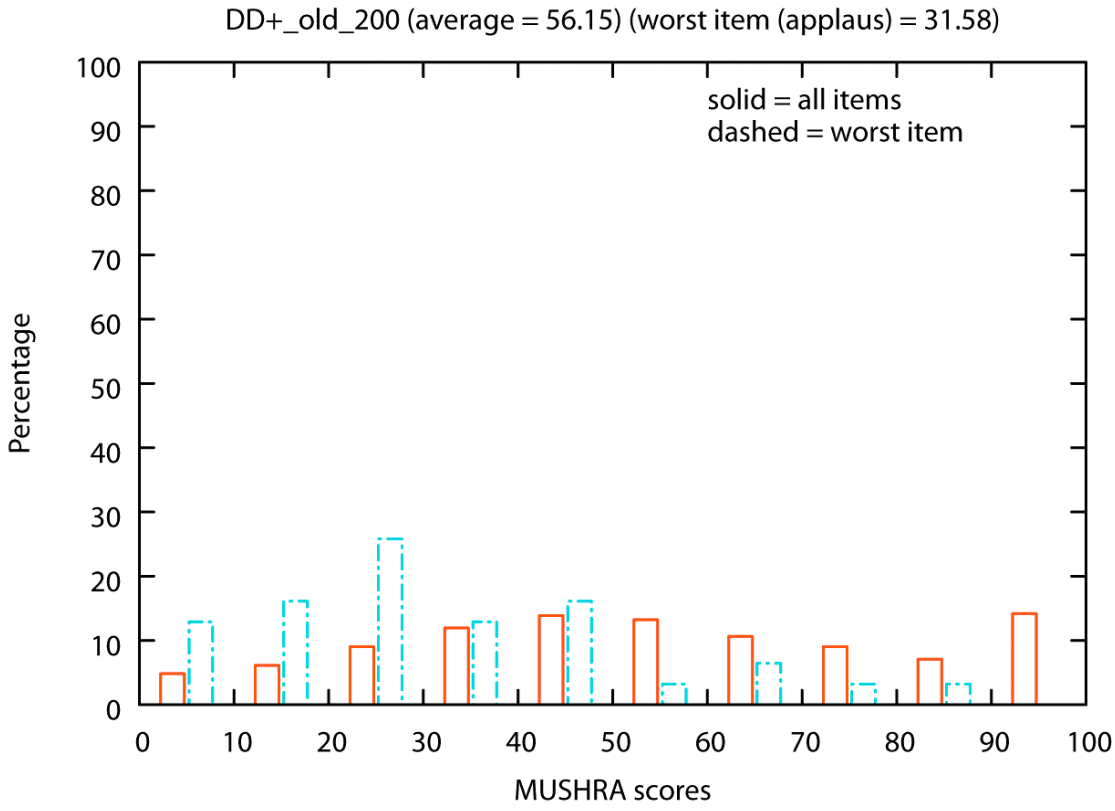

ANC_Spatial (average = 49.30) (worst item (applaus) = 32.65)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

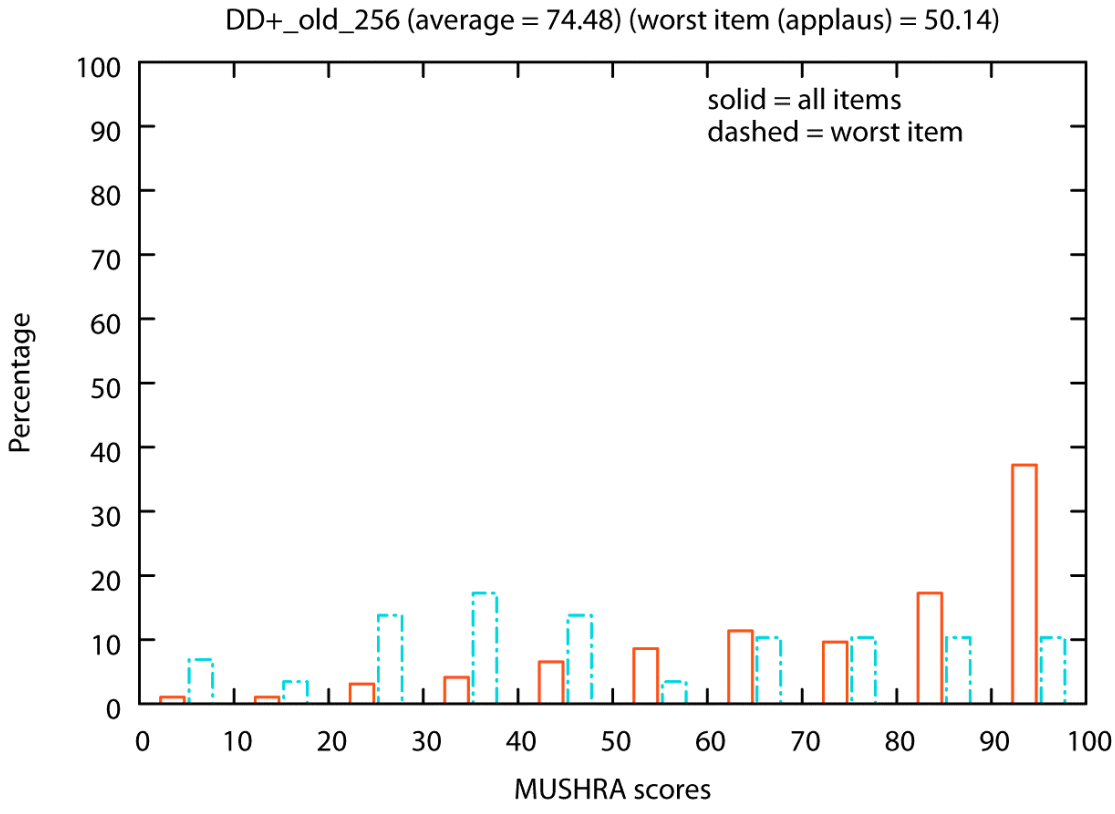

DD_448 (average = 86.47) (worst item (fleetwd) = 80.00)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

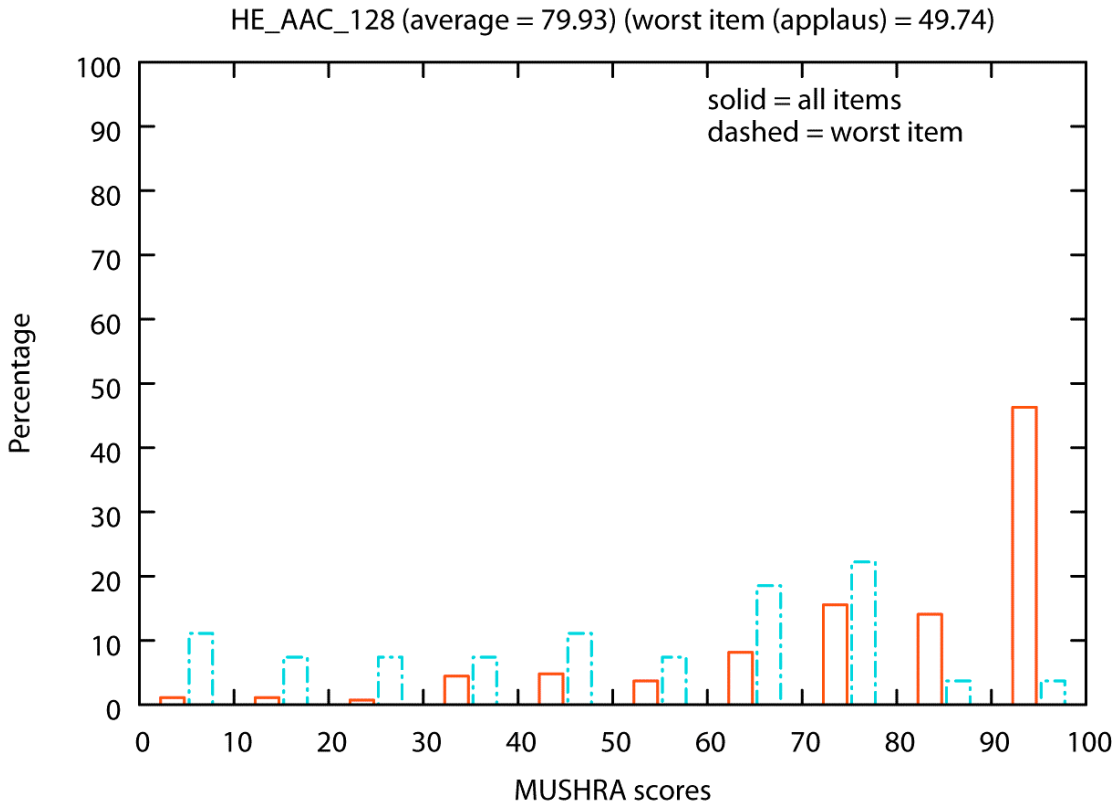

DD+256_DD+640 (average = 76.16) (worst item (applaus) = 49.41)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

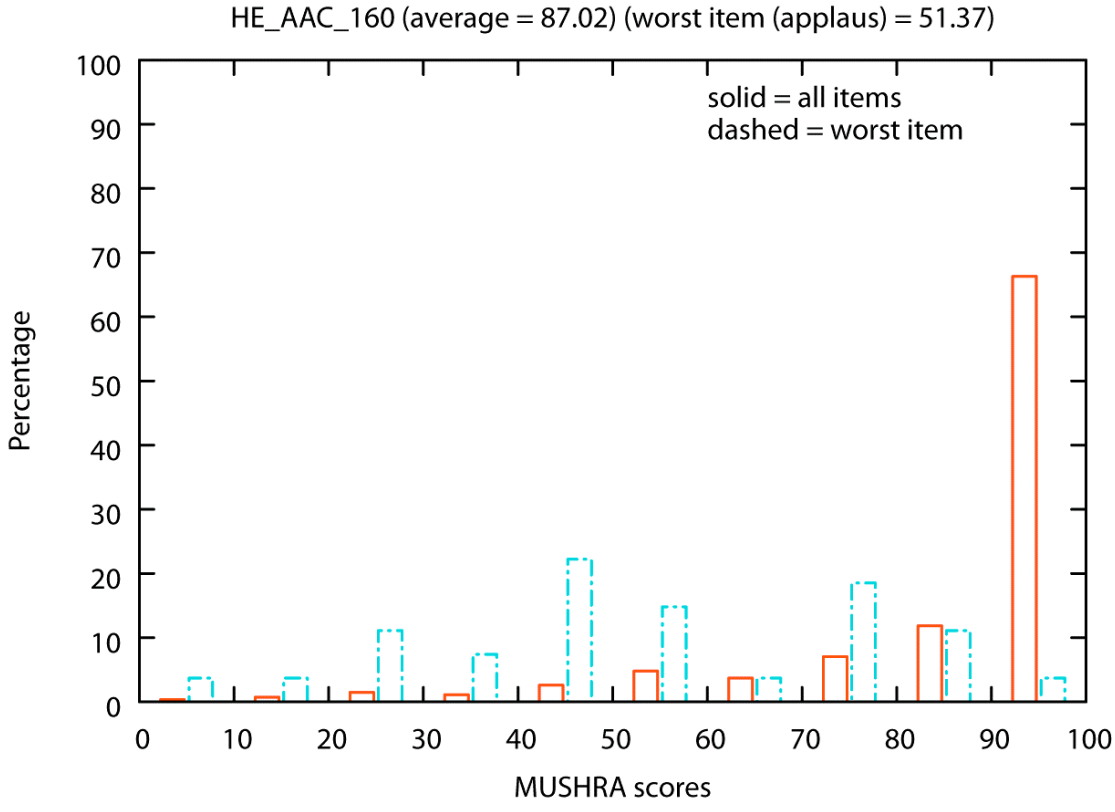

DD+_new_200 (average = 61.93) (worst item (applaus) = 40.89)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

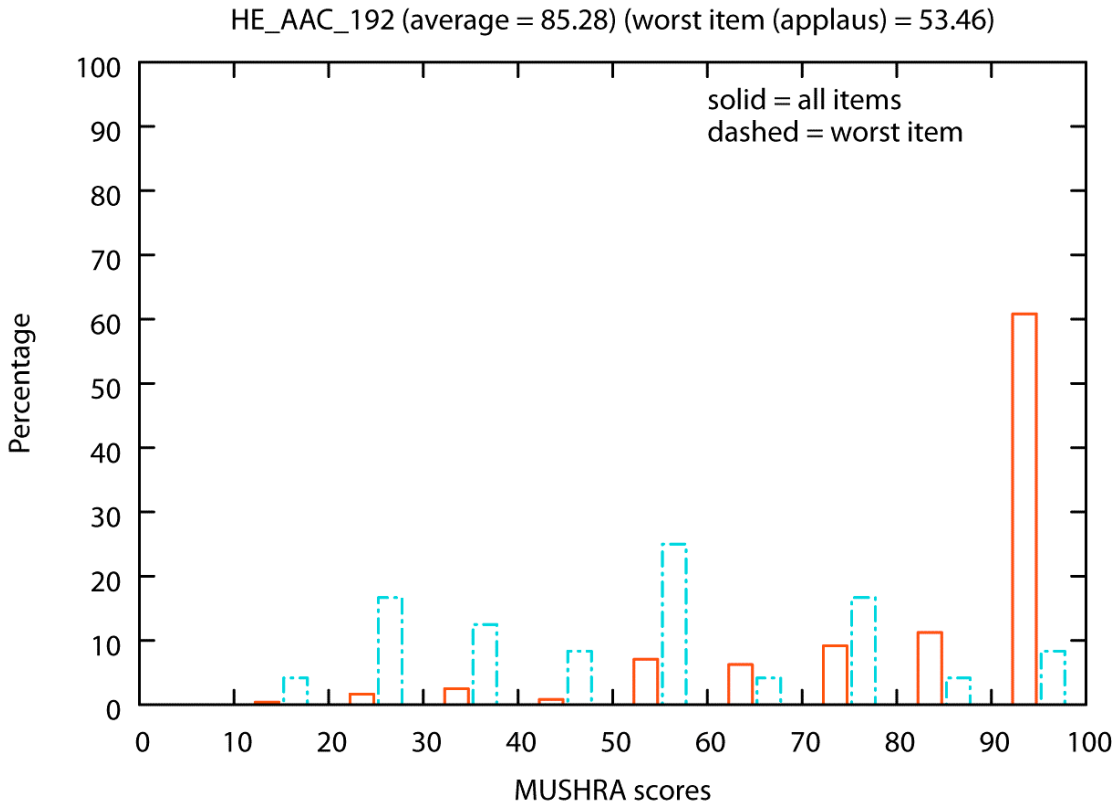

DD+_new_256 (average = 73.60) (worst item (applaus) = 46.96)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100

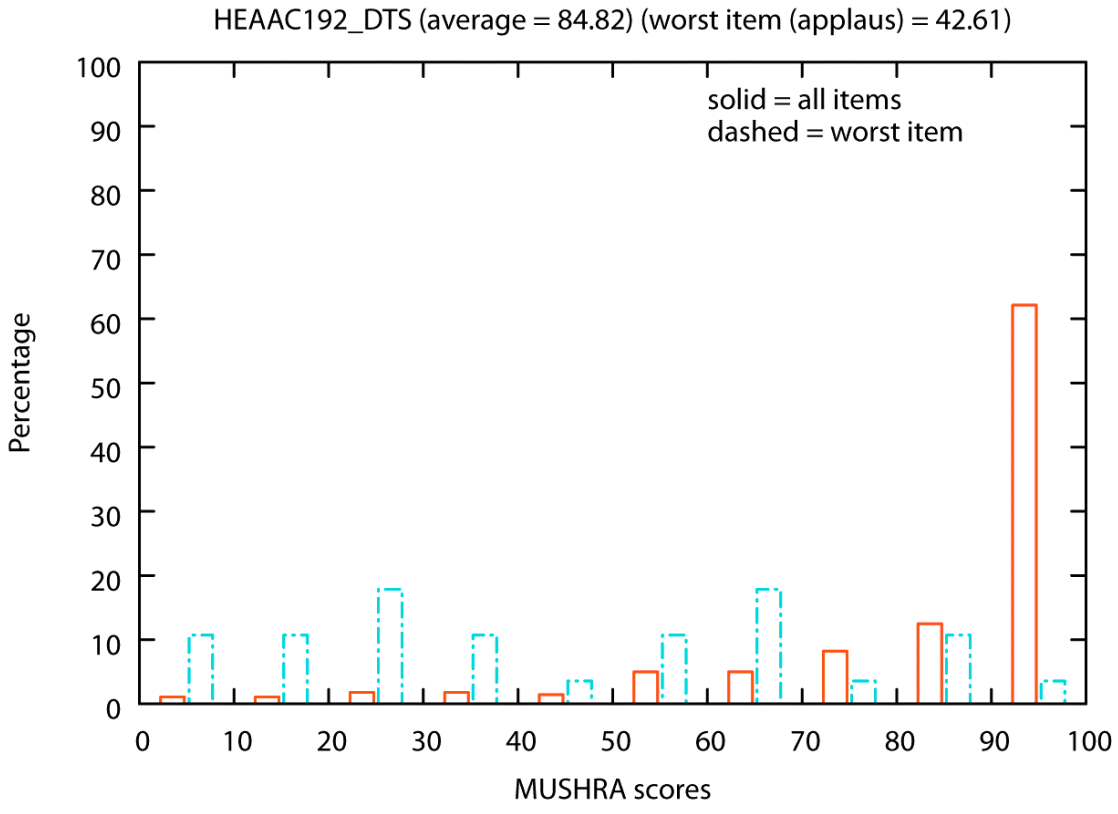

DD+_old_200 (average = 56.15) (worst item (applaus) = 31.58)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100


DD+_old_256 (average = 74.48) (worst item (applaus) = 50.14)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100


HE_AAC_128 (average = 79.93) (worst item (applaus) = 49.74)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100


HE_AAC_160 (average = 87.02) (worst item (applaus) = 51.37)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100


HE_AAC_192 (average = 85.28) (worst item (applaus) = 53.46)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100


HEAAC192_DTS (average = 84.82) (worst item (applaus) = 42.61)
solid = all items
dashed = worst item
Percentage
MUSHRA scores
0
10
20
30
40
50
60
70
80
90
100