Technical Specification Group Services and System Aspects *TSGS#28 (05)0253*
Meeting #28, Quebec, Canada, 6-9 June 2005

Source: **TSG-SA WG4**

Title: Reports of Listening / Global Analysis Laboratories related to Audio Codec Performance Characterisation (Phase 1)

Agenda Item: **7.4.3**

Presentation to: **TSG SA Meeting #28**

Presented for: **Approval**

Abstract of document:

The present document contains the reports of the Listening / Global Analysis Laboratories related to Audio Codec Performance Characterisation (Phase 1).

The Performance Characterisation tests (Phase 1) for the *Extended Adaptive Multi-Rate - Wideband* (AMR-WB+) and the *Enhanced aacPlus* codecs were performed by Coding Technologies, Dynastat, Ericsson, and FranceTelecom R&D.

The Global Analysis of the Performance Characterisation tests (Phase 1) for the *Extended Adaptive Multi-Rate - Wideband* (AMR-WB+) and the *Enhanced aacPlus* codecs was performed by Dynastat.

The present document includes information useful to network operators, service/content providers and manufacturers that will be included in TR 26.936 "*Performance characterization of the Enhanced aacPlus and Extended Adaptive Multi-Rate - Wideband (AMR-WB+) audio codecs (Release 6)*".

The approval of the reports is requested to SA#28 Plenary for ETSI to proceed to the payment of the laboratories mentioned above.

Outstanding Issues:

None.

Contentious Issues:

None.

3GPP TSG SA4 #35 **Tdoc S4-050264**
San Diego, USA, May 9-13, 2005

Source: **France Telecom**
Title: **Listening laboratory report**
Document for: **Discussion and Approval**
Agenda Item: **7**

1 Introduction

This document reports on subjective tests conducted by France Telecom for the PSS/MMS[/MBMS] Audio Codec Characterization.

France Telecom has performed the experiment 1-2 of phase 1 of audio codec characterisation described in PSS/MMS[/MBMS] Audio Codec Characterization Test Plan Version 0.5.

2 Test process

2.1 Test method

The methodology MUSHRA was used for this quality test. MUSHRA stands for MUlti Stimuli with Hidden Reference and Anchor points. This is a method dedicated to the assessment of intermediate quality.
It has been recommended at the ITU-R under the name BS.1534[1].This was developed in 1999 by the EBU Project Group B/AIM in collaboration with the ITU-R Working Party 6Q.
An important feature of this method is the inclusion of the hidden reference and bandwidth limited anchor signals. For this test, anchor points were the band-limited (3.5 and 7 kHz) reference signal.

2.2 Training phase

Each listener had a period of training, in order to get familiar with the test methodology, the use of the interface software and with the kind of quality they have to assess. This was as well an opportunity to adjust the restitution level that then remained constant during the test phase.
The training session contained 4 audio items that were part of the tests.

2.3 User Interface

The MUSHRA method has the advantage of displaying all stimuli for one test item at a given bit-rate at the same time. The subjects were therefore able to carry out any comparison between them directly as well as to assess the quality comparing to the one of the explicit reference signal.
Implementation of MUSHRA user interface from CRC (SEAQ) was used in those tests. A screenshot of one implementation of the user interface is shown in figure 1. The buttons represent all the configurations/codecs under test including the hidden reference and both the anchor signals, and the reference, which is specially displayed on the left as "REF". Above each button, with the exception of the "REF" one, a slider is used to grade the quality of the test item according to the continuous quality scale.
For each of the test items, the signals under test were randomly assigned, with a different assignment for each subject. In addition, the test items were randomised for each subject within a session to avoid sequential effects. The session files were prepared by the host lab. There was one session file per listener.
The same randomisation process was used for the training sessions : there was one training session per listener.

[1] ITU-R Recommendation BS.1534 (June 2001)/ Method for the subjective assessment of intermediate quality level of coding systems.

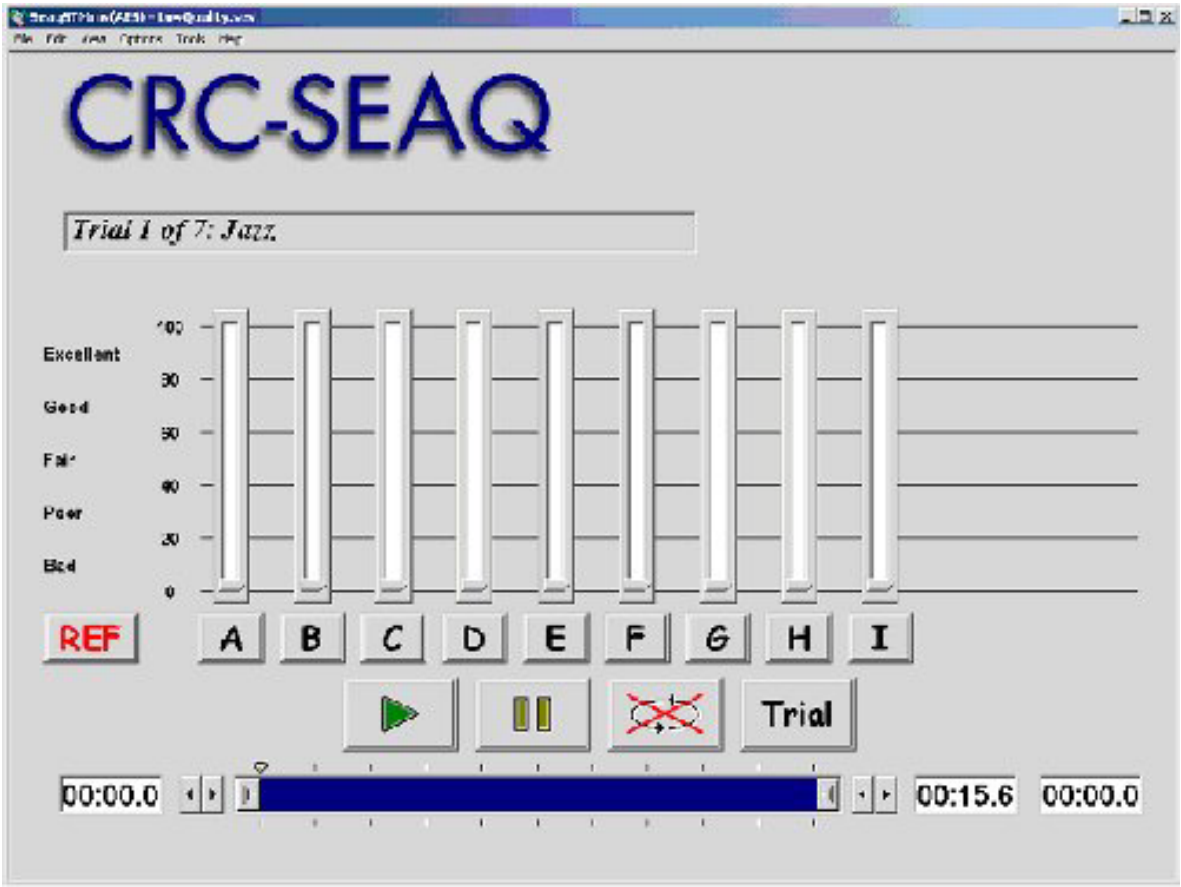


Figure 1 : MUSHRA Software

2.4 The Listening Panel

The listening panel consisted of 20 subjects, most of them experienced in audio but not only professionally involved. 5 listeners were discarded after applying the rejection process (see part 2.8). All the 15 remaining listeners were respectful regarding the listening instructions.

2.5 Tests duration

As mentioned above the test was preceded by a training period.
The training phase took about half an hour. This time was also used to describe the listening instructions and answer listeners' questions if any. If the listeners have faced difficulties in the assessment of the quality, this time was also used to explain them how to behave.
Then, one test took approximately 1 hour and 30 minutes (depending on listeners), including breaks. Every 20 minutes, the listener was asked to rest a bit by walking and breathing some fresh air.

2.6 Listening conditions

The tests were performed on the headphone STAX Signature SR-404 (open)[2] and its amplifier SRM-006t. The subjects had the possibility to set the reproduction level individually before they started the actual test (during the training phase). The subjects were then restricted from changing the reproduction level during the test.
The test items were stored on a Windows 2k workstation. The digital sound was played through the PC board Digigram VX 222 and converted by 24 bits DAC (3Dlab DAC 2000).

[2] http://www.son-video.com/Rayons/Hifi/Casques/Stax.html

The tests were run in an acoustically neutral room dedicated to such tests.

2.7 Test agenda

Test material has been received on April 5th. Raw data of test results have been sent to global analysis laboratory on April 22nd.

2.8 Rejection process

Two post-screening methods were used:

- One is based on the ability of a subject to make consistent repeated grading; and to recognize the hidden reference,
- The other relies on inconsistencies of an individual grading compared with the mean result of all subject for a given item. This was done by looking to the individual spread and to the deviation from the mean grading of all subjects. The aim of this was to get a fair assessment of the quality of the test items.

Due to the fact that "intermediate" quality is tested, a subject should be able to easily identify the coded version and therefore should be able to give a grade that is in the range of the majority of the subjects. Subjects with grades at the upper end of the scale are likely to be less critical. Subjects who have grades at the lowest end of the scale are likely to be too critical. The methods are primarily used to eliminate subjects who cannot make the appropriate discriminations.
The easiest way to measure the inconsistencies of an individual subject compared to the mean result is to calculate the correlation coefficient. This coefficient $\rho_{x,y}$ is used to determine the relationship between 2 sets of data. It is calculated as follows:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x.\sigma_y}$$

where:

$$-1 \le \rho_{x,y} \le 1 \quad \text{and} \quad Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

X : individual scoring, Y : average scoring, n : number of items . number of scores/item

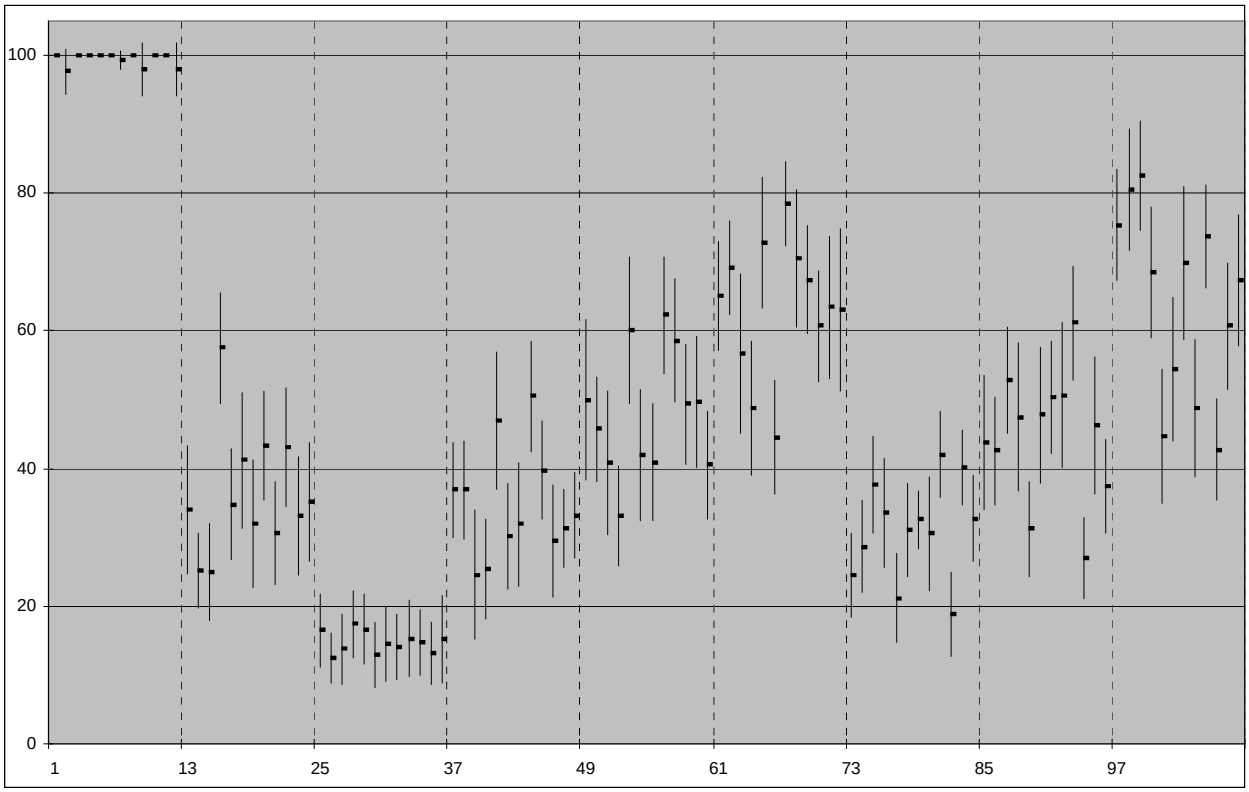
Consequently, subjects whose coefficient $\rho_{x,y}$ was below 0.83 were discarded.

2.9 Statistical analysis

The statistical analysis method described in the MUSHRA specifications was used to process the test data. The results are presented as mean grades and standards deviation.
Experience has shown that the scores obtained for different test sequences are dependent on the criticality of the test material used. Therefore, this figure have been included in this report in order to provide a more complete understanding of codec performance by presenting results for different test sequences rather than only as aggregated averages across all the test sequences used in the assessment.

3 Test results

The test results are presented below.



First column : Hidden reference

Second column : LP 7 kHz Anchor

Third column : LP 3.5 kH Anchor

Fourth column : AMR WB 14s

Fifth column : AMR WB 21s

Sixth column : AMR WB 28s

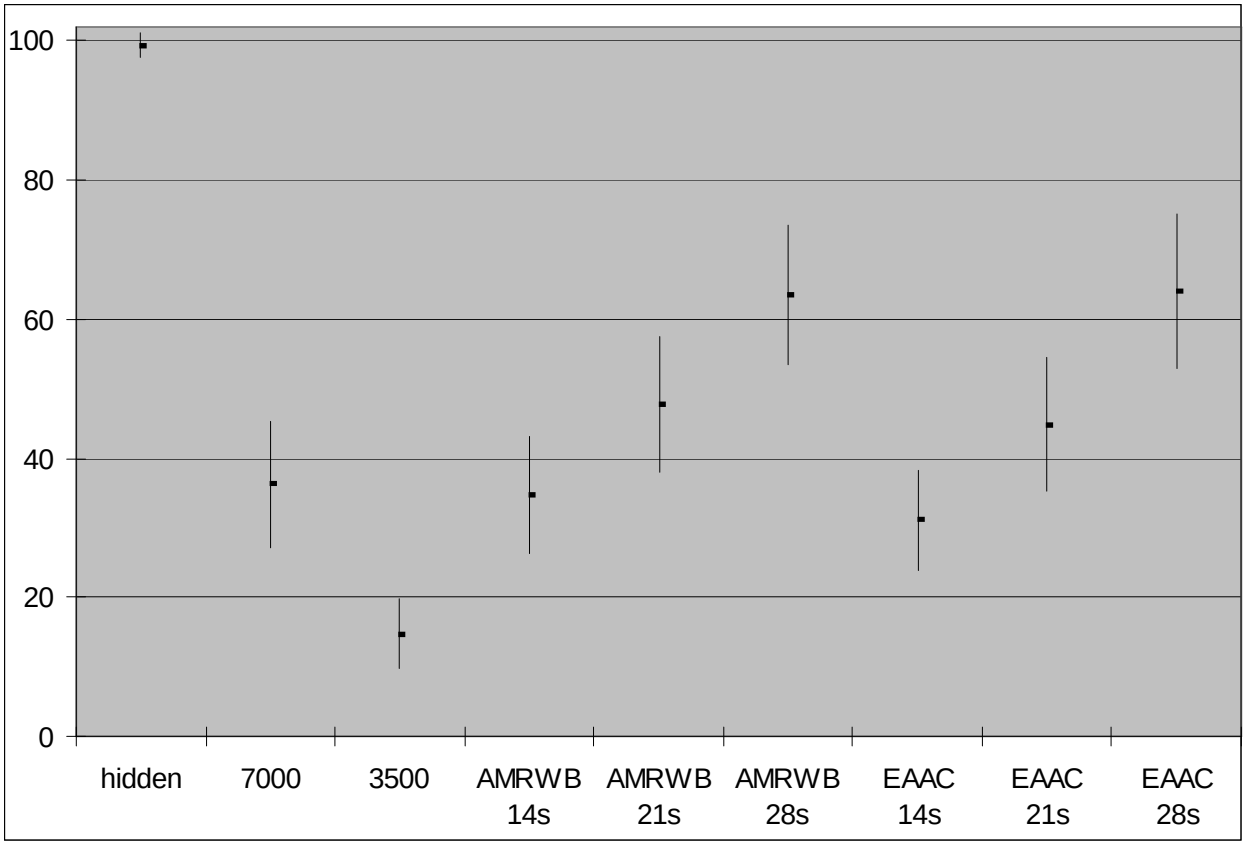
Seventh column : EAAC 14s

Eighth column : EAAC 21s

Ninth column : EAAC 28s

Sequences / column :

1- m_cl_x_2
2- m_ot_x_1
3- m_p_x_1
4- m_si_x_1
5- s_cl_2t_1
6- s_cl_2t_2
7- s_cl_mt_1
8- s_no_ft_2
9- sbm_sm_x_1
10- sbm_sm_x_2
11- som_fi_x_2
12- som_ot_x_1


100
80
60
40
20
0
hidden
7000
3500
AMRWB 14s
AMRWB 21s
AMRWB 28s
EAAC 14s
EAAC 21s
EAAC 28s

3GPP TSG SA4 #35 **Tdoc S4-050301**
San Diego, USA, May 9-13, 2005

Source: **Coding Technologies**

Title: **Coding Technologies listening report on audio codec characterization testing**

Agenda Item: **7**

Document for: **Approval**

---

1. Introduction

Coding Technologies conducted listening tests based on the work plan for the 3GPP audio codec characterization test [1]. This document describes the tests carried out at Coding Technologies and the experimental design.

2. Test cases

Coding Technologies accomplished the following tests as defined in [1]:

Phase	Exp.	Operational mode	#Codecs	# cond/codec	#Anchors	#Ref.	# Signals	#items
1	1-2	Stereo bit-rate	2	3	2	2	10	12

Table 1: Sub-experiment 1-2 carried out at Coding Technologies

3. Experimental Design

3.1 Test Method

The test procedure followed that of the “Multiple Stimulus with Hidden Reference and Anchors” (MUSHRA) [2] method for the subjective assessment of intermediate quality audio. Figure 1 shows a screenshot of the user interface of the MUSHRA implementation used at Coding Technologies. The specific MUSHRA implementation was done by Fraunhofer IIS.

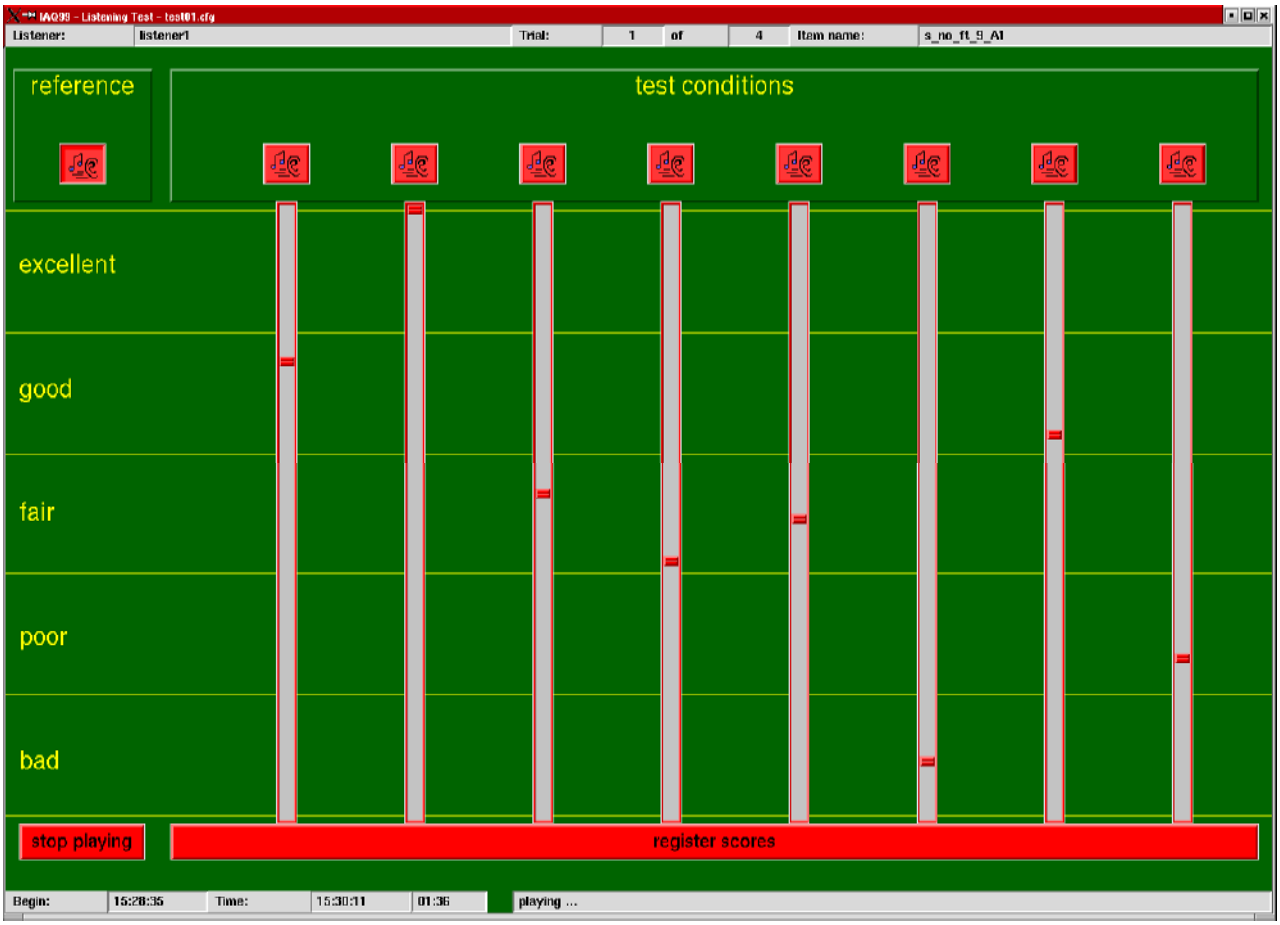


Figure 1: MUSHRA user interface

3.2 Training phase

Prior to the actual testing a training phase was carried out in which the test subjects were familiarized with testing methodology and environment. The training was done following the same MUSHRA methodology as the actual test, though limited to four trials.

3.3 Grading phase

In average the test subjects carried out the experiment in around 60-75 minutes (including training session). The test subjects were allowed to take breaks, however they were not allowed to communicate their impressions during the test with other test subjects.

3.4 Test subjects and post-screening

17 test subjects, 2 female and 15 male, aged between 26 and 47 years took part in the testing. All but two of the listeners were experienced from similar exercises, most of them with a background as audio engineers and musicians. From scanning the resulting test data for the original hidden reference it turned out that 15 of the 17 test subjects found the original hidden reference in every single experiment: Two test subjects missed the hidden reference several times and partly mixed up the anchors. Thus those two test subjects, which turned out to be the non-expert listeners, were excluded from the listening test result during post-screening.

3.5 Test schedule

The Tests started on April 7th and were finalized on April 25th. The results of the 15 test subjects after post-screening was delilvered to the Global Analysis Lab on April 25th.

3.6 Listening Laboratory

The tests were conducted in the listening rooms at Coding Technologies Nuremberg and Coding Technologies Stockholm. Both listening sites are equipped similarly: the audio data was replayed by a digital sound card to a high-class digital/analog converter and reproduced by STAX Lambda Pro headphones in a quiet acoustically controlled listening room.

4. Conclusion

The listening test at Coding Technologies has been carried according to the characterization test plan [1].

5. References

[1] SA-050188 “PSS/MMS[/MBMS] Audio Codec Characterization Test Plan Version 0.5”

[2] EBU Technical recommendation: “MUSHRA-EBU Method for Subjective Listening Tests of Intermediate Audio Quality”, Doc. B/AIM022, October 1999

Source: **Ericsson**
Title: **PSS/MMS/MBMS audio codec characterization - Test Lab report**
Document for: **Approval**
Agenda item: **7**

Executive Summary

This document contains the report of the listening lab activities for 3GPP audio codec characterization, which Ericsson has carried out by contract with ETSI. The task was to perform one listening test using the MUSHRA testing methodology, collect voting scores and deliver test results to Dynastat, according to document SP-050132 (PSS/MMS/MBMS Audio Codec Characterization Test Plan). This report specifies how in detail the task has been completed.
In conclusion, the task according to the assignment as listening test lab has been performed and no irregularities need to be reported.

Experiment

According to document SP-050132 (PSS/MMS/MBMS Audio Codec Characterization Test Plan) Ericsson performed the following experiment:

Phase	Exp.	Operational mode	#Codecs	# cond/codec	#Anchors	#Ref.	# Signals	#items
1	1-1	Mono bit-rate	2	3+1	2	2	11	12

Specifically, the experimental conditions were as follows:

Table 1: Conditions and factors for Experiment 1-1 (mono – bit rate)

Main Codec Conditions		
Codec(s)	2	AMR-WB+, E-AAC+
Use case	1	A (PSS) (except low-complexity AMR-WB+ condition)
Error Conditions	1	No errors
Mono/Stereo	1	Mono
Bit rates	3	10kbps, 16kbps, 20kbps
Low complexity-AMR-WB	1	10kbps
References		
Open Reference	1	Original signal
Hidden Reference	1	Original signal
Anchors	2	3.5 kHz and 7 kHz low-pass filtered original signal
Common Conditions		
Stimulus type		Sound item
Radio Channels	0	Clean
Number of audio items	12	
Input sampling rate		48 kHz
Number of input channels	2	Stereo input
Output sampling rate		Unspecified
Number of output channels	1	Mono
Listening Level	1	To be chosen by subject
Listeners	15	Experienced listeners
Presentation randomizations	15	One for each listener
Rating Scale	1	Continuous quality scale
Replications	2	Each sub-experiment is done in two independent test labs
Listening System	1	Binaural high-quality headphones
Listening Environment		Room Noise: Hoth Spectrum at 30dBA (as defined by ITU-T, Recommendation P.800, Annex A, section A.1.1.2.2.1 Room Noise, with table A.1 and Figure A.1)

Experiment data

According to the test plan, the listening test data (processed material) was obtained by Ericsson (host-lab) after cross-checking by CT (cross-check laboratory).

Listening panel

The listening panel was selected from experienced listeners inside Ericsson. A pre-screening procedure was used were previous performance in intermediate quality audio listening tests served as an indication of the listeners' ability to judge anchors and references in a correct way, as well as the ability to repeatedly grade in a consistent manner.
The listeners had all previous experience of audio listening tests using the Mushra methodology.

Lab setup

The listening environments were two listening labs, which both conformed to the test plan (see table 1 above). Open-back circum-aural headphones were used (Sennheiser HD650) and the listeners could individually adjust the listening level to their preference. The audio was fed from the computer to the listener using M-audio USB Duo sound cards.

Mushra test software

The Mushra software used is an in-house development. It has a similar GUI as the CRC-SEAQ software shown in the test plan. The software performs both inter-item and intra-item randomization of the test sequence, and provides a raw output of the test results into individual listener output files.

Listener instructions

See Annex 1.

Irregularities during the test

No irregularities have been encountered.

Post screening of results

An inspection of the voting scores was done after the testing. The acceptance criterion of the voting scores was the ability of the subject to rank the anchors and the hidden reference consistently in correct order. According to this criterion all votes have been found acceptable.

Testing session

Listeners were asked first to read through the listener instructions and then to start with the training phase. They were further instructed to take breaks frequently between trials depending on their own feeling in order to avoid fatigue. After the training phase and provided that there were no questions, the listeners started the grading phase.

Results

The voting scores were sent to Dynastat (GAL) in result sheets obtained from Dynastat. Results from the training phase of the test have not been delivered. In order to provide a cross-checking possibility for the result post-processing and analysis, Ericsson supplied an own basic results analysis to Dynastat, see Annex 2.

Conclusion

The task according to the assignment as listening test lab has been performed and no irregularities need to be reported.

Annex A: Listener Instruction

Instructions to be given to subjects

1 Training phase

The first step in the listening tests is to become familiar with the testing process. This phase is called a training phase and it precedes the formal evaluation or grading phase.

The purpose of the training phase is to allow you, as an evaluator, to learn how to use the test equipment and the grading scale.

In the training phase you will participate a short test similar to the one you will perform in the grading phase of the test.

No grades given during the training phase will be taken into account in the actual tests.

Please start the training phase by clicking on the "Mushra_Training" icon.

2 Grading phase

The purpose of the grading phase is to invite you to assign your grades using the quality scale. Your grades should reflect your subjective judgment of the quality level for each of the sound excerpts presented to you. Each trial will contain 10 signals to be graded. Each of the items is approximately 5 to 10 s long. You should listen to the reference and all the test conditions by clicking on the respective buttons. In a first step it is recommended to browse through all signals within each trial in order to get a coarse impression of the offered quality range. Then you may listen more carefully and start to rank the signals. You may listen to the signals in any order, any number of times. 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 "register scores" button at the bottom of the screen.

The grading scale is continuous from "excellent" to "bad". A grade of 0 corresponds to the bottom of the "bad" category, while a grade of 100 corresponds to the top of the "excellent" category.

In evaluating the sound excerpts, please note that you should not necessarily give a grade in the "bad" category to the sound excerpt with the lowest quality in the test. However one or more excerpts must be given a grade of 100 because the unprocessed reference signal is included as one of the excerpts to be graded.

You should not discuss your personal interpretation or your gradings with the other subjects at any time during the test.

Please start the grading phase by clicking on the "Mushra_Test" icon.

Annex B: Basic Statistical Result Analysis

User:	Average WB+ 10m	WB+ 10mlc	WB+ 16m	WB+ 20m	E-AAC+ 10m	E-AAC+ 16m	E-AAC+ 20m	REF
m_cl_x_2/m_cl_x_2_AMRWB+_exp_1_1_cond_10m.wav	74,67	25,87	86,13	90,47	40,40	65,27	69,07	98,67
m_ot_x_1/m_ot_x_1_AMRWB+_exp_1_1_cond_10m.wav	64,40	62,40	80,93	80,93	41,33	62,27	73,93	100,00
m_p_x_1/m_p_x_1_AMRWB+_exp_1_1_cond_10m.wav	49,00	32,80	63,87	78,87	47,13	65,67	84,13	100,00
m_si_x_1/m_si_x_1_AMRWB+_exp_1_1_cond_10m.wav	60,60	16,80	78,93	77,87	54,20	74,40	82,27	100,00
s_cl_2t_1/s_cl_2t_1_AMRWB+_exp_1_1_cond_10m.wav	56,87	40,00	81,93	83,53	18,67	46,60	64,13	100,00
s_cl_2t_2/s_cl_2t_2_AMRWB+_exp_1_1_cond_10m.wav	58,80	38,60	73,07	75,27	39,80	66,67	74,53	100,00
s_cl_mt_1/s_cl_mt_1_AMRWB+_exp_1_1_cond_10m.wav	43,60	40,33	71,67	90,53	24,93	68,47	80,00	100,00
s_no_ft_2/s_no_ft_2_AMRWB+_exp_1_1_cond_10m.wav	66,07	49,07	80,07	82,47	29,33	61,20	77,60	100,00
sbm_sm_x_1/sbm_sm_x_1_AMRWB+_exp_1_1_cond_10m.wav	62,93	38,60	82,20	88,07	43,53	67,53	83,33	100,00
sbm_sm_x_2/sbm_sm_x_2_AMRWB+_exp_1_1_cond_10m.wav	57,13	46,67	79,33	81,67	22,47	47,67	59,73	100,00
som_fi_x_2/som_fi_x_2_AMRWB+_exp_1_1_cond_10m.wav	64,20	34,60	70,13	77,33	43,60	71,47	77,87	100,00
som_ot_x_1/som_ot_x_1_AMRWB+_exp_1_1_cond_10m.wav	58,00	35,20	67,73	80,33	26,73	65,67	75,73	100,00

Average	WB+ 10m	WB+ 10mlc	WB+ 16m	WB+ 20m	E-AAC+ 10m	E-AAC+ 16m	E-AAC+ 20m	REF
M	62,17	34,47	77,47	82,03	45,77	66,90	77,35	99,67
S	56,33	42,00	76,68	82,95	28,18	60,73	74,07	100,00
SbM	60,03	42,63	80,77	84,87	33,00	57,60	71,53	100,00
SoM	61,10	34,90	68,93	78,83	35,17	68,57	76,80	100,00
Total	59,69	38,41	76,33	82,28	36,01	63,57	75,19	99,89

Min								
M	49,00	16,80	63,87	77,87	40,40	62,27	69,07	98,67
S	43,60	38,60	71,67	75,27	18,67	46,60	64,13	100,00
SbM	57,13	38,60	79,33	81,67	22,47	47,67	59,73	100,00
SoM	58,00	34,60	67,73	77,33	26,73	65,67	75,73	100,00
Total	43,60	16,80	63,87	75,27	18,67	46,60	59,73	98,67

Max								
M	74,67	62,40	86,13	90,47	54,20	74,40	84,13	100,00
S	66,07	49,07	81,93	90,53	39,80	68,47	80,00	100,00
SbM	62,93	46,67	82,20	88,07	43,53	67,53	83,33	100,00
SoM	64,20	35,20	70,13	80,33	43,60	71,47	77,87	100,00
Total	74,67	62,40	86,13	90,53	54,20	74,40	84,13	100,00

3GPP TSG-SA4 Meeting #30 ***Tdoc S4-050351***
San Diego, California, 9-13 May 2005

Source: **Dynastat**[1]

Title: **Dynastat Listening Laboratory Report for the 3GPP Audio Codec Characterization Test**

Agenda item: **7**

---

1. Introduction

Dynastat performed a single listening test, Experiment 1-1 (*Monaural/Bit-rate*), in accordance with the test plan for the characterization of the 3GPP Audio Codecs [1]. The test plan specified the experimental design, methodology, and test procedures for conducting the test. Dynastat received the processed files from the Host Laboratory, conducted the test, and delivered raw subjective data to the Global Analysis Laboratory. In conducting the test, Dynastat complied with all of the test methods, procedures, and schedule specified in the test plan.

2. Experimental Design

2.1. Test Method

The test procedure followed that of the "Multiple Stimulus with Hidden Reference and Anchors" or MUSHRA method [2] for the subjective assessment of intermediate audio quality.

Subjects were presented with a series of trials, each corresponding to a different item from the set of audio items selected for the tests. In each trial, the subject was presented with the open reference version of the item as well as a set of test signals to be graded.

The test signals consisted of two codecs at different bit rates, two anchors, and a hidden copy of the open reference for a total of ten test signals. The anchors were bandwidth-limited versions of the unprocessed reference signal and were defined as 3.5 kHz Low-pass and 7.0 kHz Low-pass by the test plan [1]. The test conditions included the codecs AMR-WB+ and EAAC+ at bit rates of 10, 16, and 20kbps. A low complexity version of the AMR-WB+ codec at 10kbps was also included.

An in-house MUSHRA presentation and data collection interface program was used for this effort. A sample MUSHRA presentation screen for the Dynastat proprietary interface is shown in Figure 1. The Dynastat MUSHRA interface program complies with the specifications contained in the MUSHRA standard.

The open reference was shown on one button followed spatially by buttons, labelled *A* through *J*, for the signals to be graded. The grading scale varied from 0 to 100 in unit steps and grades were recorded by adjusting the slider associated with each button. The MUSHRA presentation program allowed clean switching among all of the signals even during play-back. Both the order of presentation of the trials and the allocation of the signals to the buttons (A through J) were independently randomized for each subject.

---

[1] Alan Sharpley
Dynastat, Inc Email: sharpley@dynastat.com
2704 Rio Grande Phone: +1-512-476-4797
Austin, Texas, USA 78705 FAX: +1-472-2883

Dynastat received the processed audio files from the Host Lab. The corpus of audio materials included 12 test items, four items for each of three classes of audio content - Speech-only, Music-only, Mixed speech and music content. Also included were four training items - one Speech-only, one Music-only, and two Mixed-content.

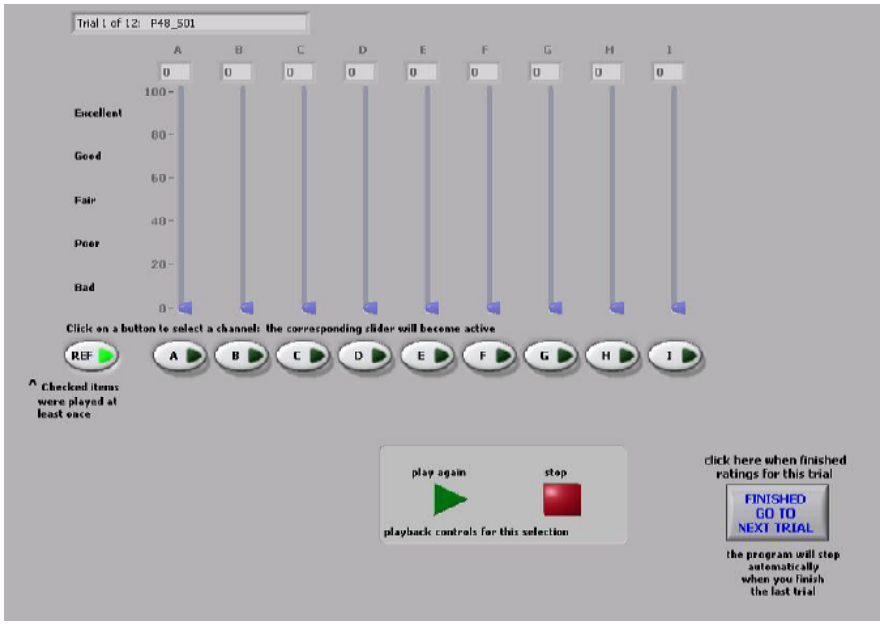


Figure 1: Sample Presentation Screen for the Dynastat MUSHRA Interface Program.

2.2. Training phase

Prior to the actual grading of the test signals, a training phase was conducted in which the test subjects were familiarized with the testing methodology and testing environment. The training phase adhered to the same MUSHRA methodology as the grading phase, but was limited to four trials. Subjects were provided with written instructions prior to participating in any experiment.

2.3. Grading phase

Each listener received a different randomized presentation sequence of items and signals within items. The grading phase was preceded by the training phase and was separated from the training phase by a forced rest break. In order to mitigate the effects of fatigue, subjects were required to take two additional rest breaks during the test session -- one after each group of four test items. In addition, the tests were self-paced so subjects could take additional breaks if they wanted.

2.4. Listening panels

Twenty subjects were selected from Dynastat's pool of expert listeners for the MUSHRA experiment. All were under the age of 45 and had previous experience with the MUSHRA task or with other critical listening subjective experiments. Fifteen listeners passed the Dynastat performance criteria for inclusion in the experiment based on statistical measures of self-consistency as well as the ability to track the embedded anchors and hidden reference. Ten of the subjects were male, five were female. Average age of the fifteen listeners in the listening panel was 28.5 years.

2.5. Listening environment

The tests were performed in individual sound isolation booths at Dynastat in Austin, Texas, USA. Each booth met the requirements specified in the test plan [1]. The audio materials were presented over Sennheiser HD-600 open-back circum-aural headphones. The audio level was set by the subject at the beginning of the training phase and further level adjustments were not permitted during the test session. The audio files were stored on a Windows 2000 workstation which had a digital interface board (Lynx One Studio). This board was connected to an external Lucid DA9624 digital-to-analog converter and presented over the headphones.

3. Results

Table 1 shows summary results (Means, Standard Deviations, and 95% Confidence Intervals) for the conditions evaluated in MUSHRA Exp. 1-1 (Monaural/Bit-rate). All results are based on 180 MUSHRA votes (15 subjects x 12 items). Figure 2 presents the summary scores graphically.

Table 1. Summary results for MUSHRA Experiment 1-1 (Monaural / Bit-rate)

Condition	Bit-rate	Mean	StdDev	95% CI
AMR-WB+	10k (lc)	43.72	21.62	3.16
	10k	63.58	19.39	2.83
	16k	78.13	17.65	2.58
	20k	84.46	16.86	2.46
EAAC+	10k	42.76	22.81	3.33
	16k	67.11	19.03	2.78
	20k	80.46	14.99	2.19
References	3.5k lp	36.74	19.25	2.81
	7.0k lp	71.98	18.81	2.75
	hid ref	99.92	0.62	0.09

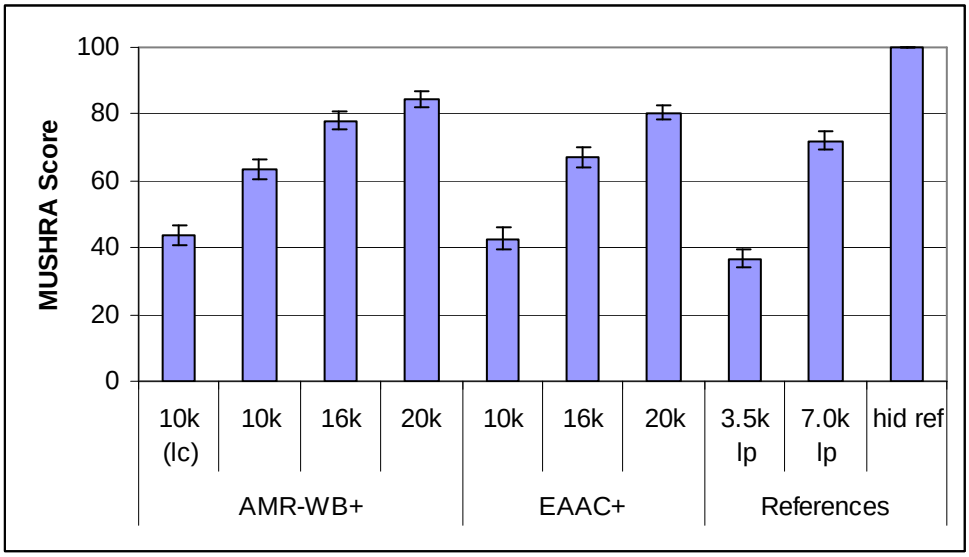


Fig 2. Summary results for MUSHRA Experiment 1-1 (Monaural / Bit-rate)

Table 2 shows MUSHRA mean scores for the test conditions by class of audio signal. Figure 3 illustrates those results graphically. Results shown in Table 2 and Fig. 3 are based on 60 MUSHRA votes (15 subjects x 4 items).

Table 2. MUSHRA Results for Test and Reference Conditions by Audio Content

Condition	Bit-rate	Speech	Music	Mixed
AMR-WB+	10k (lc)	43.17	44.63	43.37
	10k	62.08	70.48	58.17
	16k	82.30	77.47	74.63
	20k	87.82	84.27	81.30
EAAC+	10k	40.57	48.85	38.85
	16k	64.98	69.35	66.98
	20k	78.77	80.70	81.90
References	3.5k lp	39.15	37.30	33.78
	7.0k lp	75.02	73.28	67.65
	hid ref	99.97	99.88	99.90

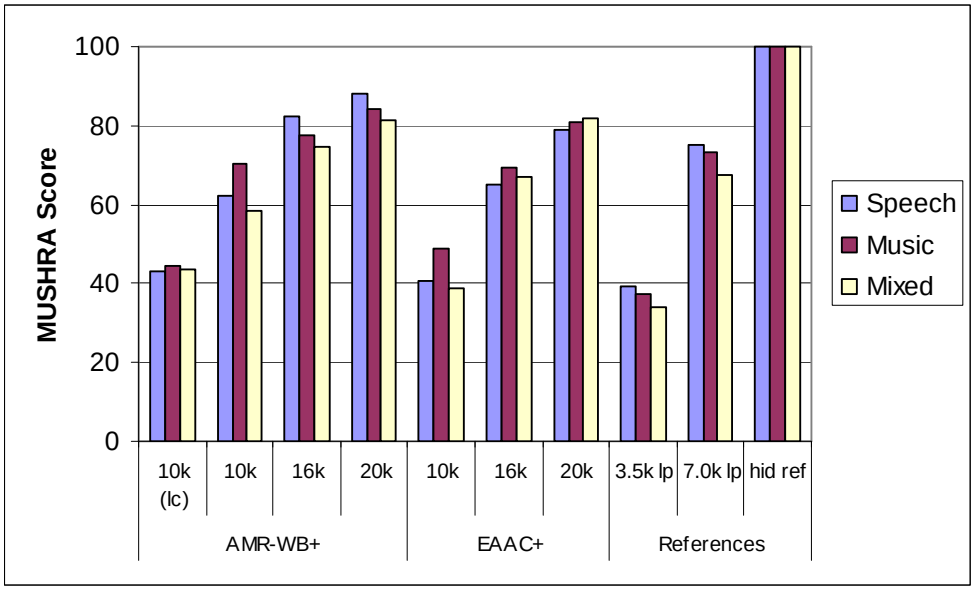


Fig. 3 MUSHRA Results for Test and Reference Conditions by Audio Content.

4. References

[1] S4-050188-PSS/MMS[/MBMS] Audio Codec Characterization Test Plan Version 0.5, Feb.2005.

[2] EBU Technical recommendation: MUSHRA-EBU Method for Subjective Listening Tests of Intermediate Audio Quality, Doc. B/AIM022, Oct.1999.

3GPP TSG-SA4 Meeting #35 ***Tdoc S4-050428***
San Diego, California, 9-13 May 2005

Source: **Dynastat**[1]

Title: **Global Analysis Laboratory Report for Phase-1 of the 3GPP Audio Codec Characterization Test for PSS-MMS-MBMS**

Agenda item: **7**

---

1. Introduction

This document comprises the final report for the Phase I activities of the Global Analysis Laboratory for the Characterization of the 3GPP Audio Codecs. It summarizes the results and analyses from Phase 1 of the Characterization Test. Phase 1 included four listening tests evaluating the subjective performance of the two audio codecs at different bit rates.

2. Organization of the Characterization Test

The Characterization Test Plan [1] specified the subjective listening tests to characterize the performance of the two audio codecs, *3GPP Enhanced aacPlus* (EAAC+) and *Extended AMR-WB* (AMR-WB+), selected by 3GPP for standardization for PSS-MMS-MBMS. The test plan specified subjective tests using the "Multiple Stimulus with Hidden Reference and Anchors" or MUSHRA test method [2] for the subjective assessment of intermediate audio quality. The MUSHRA experiments were subdivided into two phases of testing:

- Phase 1: Characterization of the two selected codecs across bit rates
  - Experiment 1-1 – Mono with bit rates 10kbps, ≈16kbps, and 20kbps
  - Experiment 1-2 – Stereo with bit rates 14kbps, 21kbps, and 28kbps

- Phase 2: Characterization of the two selected codecs across error conditions (***tbd***)

For the Phase 1 series of tests, the test plan specified that each experiment would be conducted by two Listening Labs (LL). The analyses and results of the data from the Phase 1 tests are included in this document.

Table 1 shows the test and reference conditions specified in test plan for the two Phase 1 MUSHRA experiments. Experiment 1-1 involved the two audio codecs in mono mode at three bit-rates, 10k, 16k, and 20kbps. For the AMR-WB+ codec a low complexity version at 10kbps, designated *10k(lc)*, was also included in Exp.1-1. Experiment 1-2 involved the two audio codecs in stereo mode at three bit-rates, 14k, 21k, and 28kbps. Both experiments also included three standard MUSHRA reference conditions in the appropriate Mono and Stereo mode -- 3.5k low-pass, 7.0k low-pass, and the hidden reference. Also included in Table 1 are the actual values of bit-rate for the two codecs for each of the two experiments. Dynastat (*Dyna*) and Ericsson (*Eric*) conducted the Exp.1-1 listening tests. France Telecom R&D (*FTRD*) and Coding Technologies (*CodT*) conducted the listening tests for Exp.1-2.

As specified in the test plan, each of the four MUSHRA tests involved the same 12 audio items where the audio file was processed through each test and reference condition. The audio items were selected to represent three classes of Audio Content – Music, Speech, and Mixed Music/Speech Content. The

[1] Alan Sharpley
Dynastat, Inc Email: sharpley@dynastat.com
2704 Rio Grande Phone: +1-512-476-4797
Austin, Texas, USA 78705 FAX: +1-472-2883

Mixed Content class was further sub-classified into Speech Over Music and Speech Between Music. Among the 12 test audio items, the distribution of Audio Content was as follows:

- Speech content – 4 items
- Music content – 4 items
- Mixed content – 4 items
  - Speech Over Music – 2 items
  - Speech Between Music – 2 items

The test plan required each LL to deliver raw voting data for each of 15 expert listeners for each of the 12 test items. The GAL provided each LL with an Excel spreadsheet for delivery of the raw voting data. Each LL delivered raw MUSHRA voting data to the GAL for 15 expert listeners within the deadline prescribed by the test plan.

Table 1. Listening Labs and Test/Reference Conditions involved in the Phase 1 Experiments

Exp.1-1 (Mono)			Exp.1-2 (Stereo)		
Test Conditions	Bit-rate	Actual	Test Conditions	Bit-rate	Actual
AMR-WB+	10k	9.75k	AMR-WB+	14k	14.25k
AMR-WB+	16k	15.2k	AMR-WB+	21k	20k
AMR-WB+	20k	19k	AMR-WB+	28k	27k
AMR-WB+ (lc)	10k	9.75k	EAAC+	14k	16k
EAAC+	10k	10k	EAAC+	21k	21k
EAAC+	16k	16k	EAAC+	28k	28k
EAAC+	20k	20k			
Reference Conditions			Reference Conditions		
3.5k low pass anchor			3.5k low pass anchor		
7.0k low pass anchor			7.0k low pass anchor		
Hidden reference			Hidden reference		
Listening Labs			Listening Labs		
Dynastat (Dyna)			Coding Technoligies (CodT)		
Ericsson (Eric)			France Telecom R&D (FTRD)		

3. Overall Results

Table 2 shows summary results for Exp. 1-1. The results include Means, Standard Deviations (SD) and 95% Confidence Intervals (CI-95) for each test and reference condition and for each LL. The statistics shown in the table are based on 180 votes (15 subjects x 12 test items). Figure 1 illustrates the results from Table 2.

Table 2. Summary Results for Experiment 1-1 – Mono / Bit-rate

Exp.1-1 Mono/Bit-rate		Dyna			Eric		
		Mean	StdDev	95%CI	Mean	StdDev	95%CI
AMR-WB+	10k (lc)	43.722	21.621	3.159	38.411	18.213	2.661
	10k	63.578	19.386	2.832	59.689	17.622	2.574
	16k	78.133	17.648	2.578	76.333	15.455	2.258
	20k	84.461	16.860	2.463	82.278	14.520	2.121
EAAC+	10k	42.756	22.805	3.332	36.011	18.323	2.677
	16k	67.106	19.033	2.781	63.572	17.277	2.524
	20k	80.456	14.992	2.190	75.194	16.048	2.344
References	lp 3.5k	36.744	19.250	2.812	40.861	13.905	2.031
	lp 7.0k	71.983	18.813	2.748	69.256	12.612	1.843
	hid ref	99.917	0.624	0.091	99.889	1.491	0.218

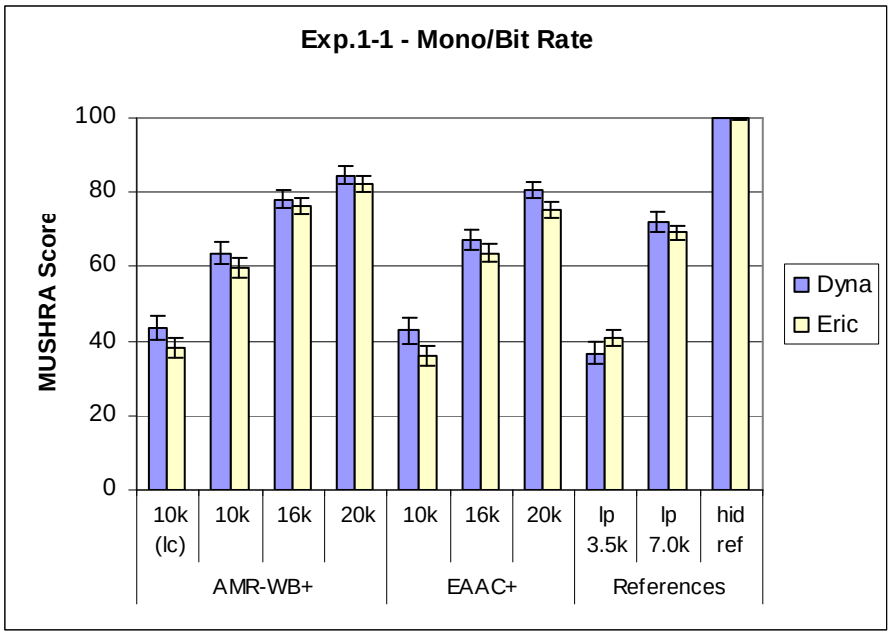


Fig.1. MUSHRA Scores for Exp. 1-1 (Mono) -- Codec x Bit-rate by LL

Table 3 shows summary results for Exp. 1-2. As in Table 2, the statistics in Table 3 are based on 180 votes (15 subjects x 12 test items). Figure 2 illustrates the results shown in Table 3.

Table 3. Summary Results for Experiment 1-2 – Stereo / Bit-rate

Exp.1-2 Stereo/Bit-rate		FTRD			CodT		
		Mean	StdDev	95%CI	Mean	Stdev	95% CI
AMR-WB+	14k	34.744	16.950	2.476	43.078	15.406	2.251
	21k	47.739	19.789	2.891	52.372	14.805	2.163
	28k	63.383	20.075	2.933	62.917	15.543	2.271
EAAC+	14k	31.133	14.410	2.105	43.350	12.934	1.890
	21k	44.872	19.449	2.841	54.839	13.676	1.998
	28k	64.083	22.103	3.229	66.639	14.682	2.145
References	lp 3.5k	14.711	9.939	1.452	24.050	9.048	1.322
	lp 7.0k	36.244	18.208	2.660	50.189	14.027	2.049
	hid ref	99.417	3.747	0.547	100.000	0.000	0.000

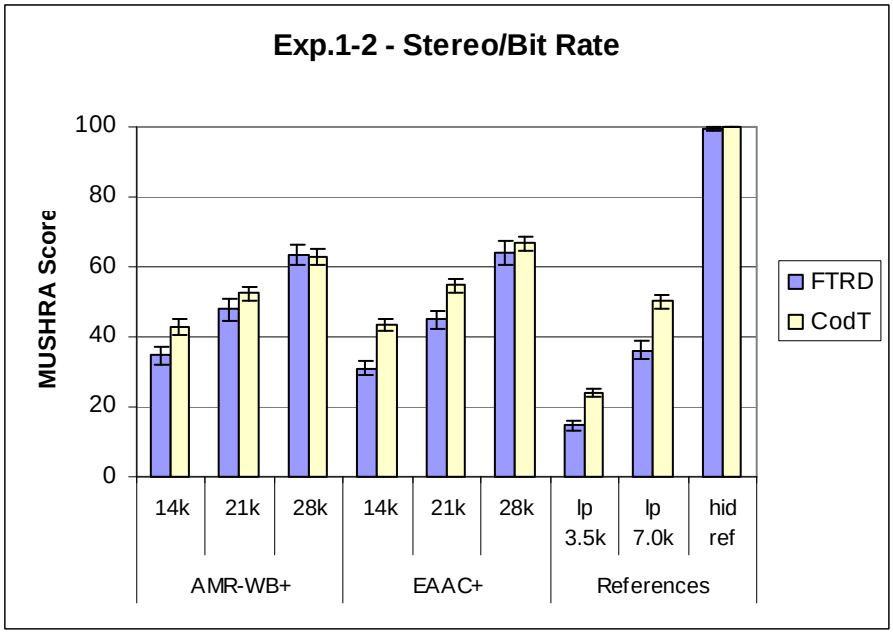


Fig. 2. MUSHRA Scores for Exp. 1-2 (Stereo) -- Codec x Bit-rate by LL

4. Lab Dependency

One of the primary goals of the Phase 1 listening tests was to determine if there were significant difference among Listening Labs in the MUSHRA scores for the audio codecs, i.e., was Lab dependency a significant factor. An examination of Figs. 1 and 2 suggests that the results from the two LL's are highly correlated in both of the Phase 1 experiments, confirmed by the computed correlation coefficients, r_{LL} = 0.989 for Exp.1-1 and 0.988 for Exp.1-2. The overall difference in MUSHRA scores across LL's is relatively small for Exp.1-1 (Diff.=2.8, $\text{Mean}_{\mathbf{Dyna}}$ = 66.9, $\text{Mean}_{\mathbf{Eric}}$ = 64.1) and somewhat larger for Exp.1-2 (Diff.=6.8, $\text{Mean}_{\mathbf{FTRD}}$ = 48.5, $\text{Mean}_{\mathbf{CodT}}$ = 55.3).

To test whether there was significant Lab Dependency for the MUSHRA results, Analysis of Variance (ANOVA) was conducted for each of the two experiments, Mono and Stereo. For both experiments the ANOVA used only the six conditions representing the two codecs at the three bit rates. The ANOVA was a nested factorial design with fixed factors *Codecs* (AMR-WB+ vs. EAAC+) and *Bit-rates* (10k, 16k, 20k) and random factor *Votes* (15 subjects x 12 items). Furthermore, the *Votes* factor was partitioned into the fixed factors *Labs* and random factor *Votes within Labs*.[1] Table 4 shows the results of the ANOVA for Exp.1-1. The critical effects to test *Lab Dependency* effects are the main effect for *Labs* and the interaction effects for *Codecs x Labs* and *Codecs x Bit-rates x Labs*.

The only significant Lab Dependency effect in Table 4 (Exp.1-1, Mono) is for the main effect for *Labs*. The significant F-ratio for *Labs* ($F = 9.61$, $p<.05$) indicates that the overall mean for the *Dyna* Lab (69.1) was significantly higher than the overall mean for the *Eric* Lab (65.5). Furthermore, the interaction effects, *Codecs x Labs* and *Codecs x Bit-rates x Labs*, are not significant indicating that the

[1] The random factor *Votes* actually contains two nested factors, *Labs* and *Audio Content*. For the analyses presented in this section the systematic effects of *Audio Content* were removed from the *Votes* factor. With the effects of Audio Content removed, each of the 12 items for each of the 15 subjects in each test can be considered as independent votes for purposes of the ANOVA.

patterns of scores for the two codecs and for the two sets of codecs by bit-rates were equivalent across LL's. For Exp.1-1 there appears to be little evidence of a *Lab Dependency* effect -- the difference in scores for the two LL's is a constant.

The other significant effects in the AVOVA are those for *Codecs* where AMR-WB+ (74.08) scored significantly higher than EAAC+ (60.05) and for *Bit-rate* where the scores for the three bit rates, averaged over codecs and labs, were significantly different (50.51, 71.29, 80.60).

Table 4. Results of Lab Dependency ANOVA for Exp. 1-1 (Mono/Bit-rate)

Exp.1-1 - Cod x BR x Lab (AC rem.)	df		Sum of Squares		Mean Square		F-Ratio	Prob.
Codecs	1		94511		94511.0		282.54	0.0000
BitRates	2		341695		170847.5		872.32	0.0000
Votes	359		313553		873.4			
Labs		1		8194		8194.0	9.61	0.0021
Votes w/n Labs		358		305359		853.0		
Codecsx BitRates	2		25604		12802.0		97.41	0.0000
Codecs x Votes	359		120636		336.0			
Codecs x Labs		1		883		883.0	2.64	0.1051
Codecs x Votes w/n Labs		358		119753		334.5		
BitRates x Votes	718		140899		196.2			
BitRates x Labs		2		668		334.0	1.71	0.1816
BitRates x Votes w/n Labs		716		140231		195.9		
Codecs x BitRates x Votes	718		94143		131.1			
Codecs x BitRates x Labs		2		42		21.0	0.16	0.8522
Codecs x BitRates x Votes w/n Labs		716		94101		131.4		
Total	2159		1131041					

Table 5 shows the results of the ANOVA for Exp.1-2. Again, the main effect for *Labs* in Table 5 is significant ($F = 28.11$, $p<.05$) indicating that the mean for the *CodT* Lab (53.87) was significantly higher than the mean for the *FTRD* Lab (47.66). Furthermore, the interaction effect for *Codecs x Labs* is also significant ($F = 5.86$, $p<.05$) indicating that the patterns of scores for the two codecs were significantly different across LL's. For Exp.1-2 there is evidence of a *Lab Dependency* effect.

There is no significant effect for *Codecs* in Exp. 1-2 – AMR-WB+ (50.71), EAAC+ (50.82). The other significant main effect in the AVOVA for Exp. 1-2 is for *Bit-rate* where the scores for the three bit rates were significantly different (38.08, 49.56, 64.26). The three significant interactions, *Codecs x Bit-rates*, *Codecs x Labs*, and *Bit-rates x Labs*, are illustrated in Figs. 3a, 3b, and 3c, respectively.

Table 5. Results of Lab Dependency ANOVA for Exp. 1-2 (Stereo/Bit-rate)

Exp.1-2 - Cod x BR x Lab (AC rem.)	df		Sum of Squares		Mean Square		F-Ratio	Prob.
Codecs	1		7		7.0		0.02	0.8876
BitRates	2		247426		123712.8		689.68	0.0000
Votes	359		286264		797.4			
Labs		1		20840		20839.5	28.11	0.0000
Votes w/n Labs		358		265425		741.4		
Codecsx BitRates	2		1385		692.3		6.16	0.0022
Codecs x Votes	359	1	139468		388.5			
Codecs x Labs		1		2247		2247.0	5.86	0.0160
Codecs x Votes w/n Labs		358		137221		383.3		
BitRates x Votes	718		136387		190.0			
BitRates x Labs		2		7953		3976.5	22.17	0.0000
BitRates x Votes w/n Labs		716		128434		179.4		
Codecs x BitRates x Votes	718		80577		112.2			
Codecs x BitRates x Labs		2		120		60.0	0.53	0.5888
Codecs x BitRates x Votes w/n Labs		716		80457		112.4		
Total	2159		891512					

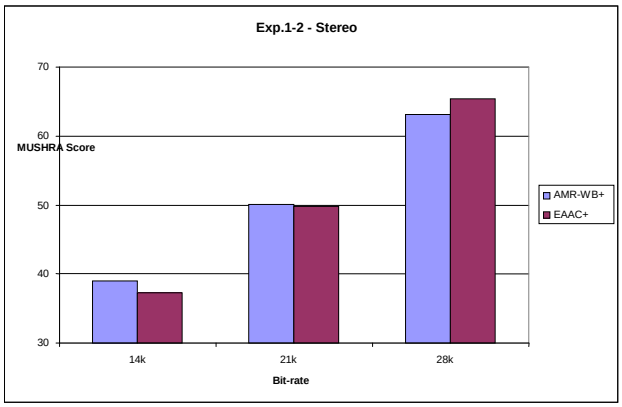


Fig.3a Interaction of Codecs x Bit Rates.

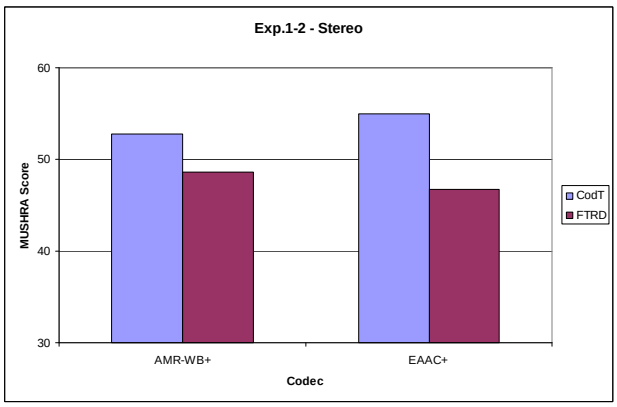


Fig.3b Interaction of Codecs x Labs.

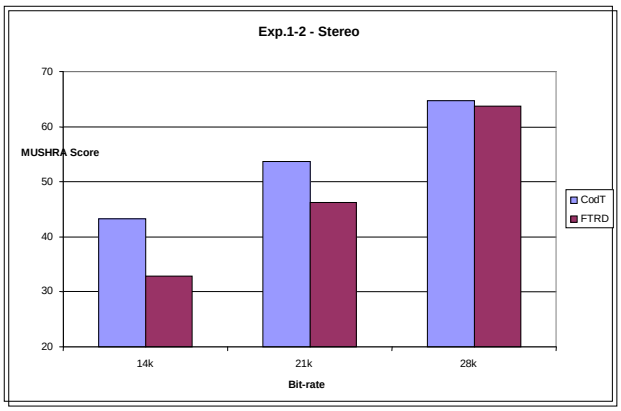


Fig.3c Interaction of Bit Rates x Labs.

5. Audio content

The twelve audio items involved in each experiment were chosen to represent three classes of Audio Content. Four of the items were classified as *Music* only, four as *Speech* only, and four as *Mixed* content – speech and music. For the analyses and figures presented in this section, the effects of the nested factor *Labs* has been removed in order to evaluate the effects of *Audio Content* independent of the effects of Listening Labs, i.e., the unconfounded effects of Audio Context.

Table 6 shows the ANOVA for Exp.1-1 for the effects of the factors *Codecs*, *Bit-rates*, and *Votes* with the fixed factor *Audio Content* nested within *Votes*. The significance of the effects for *Codecs*, *Bit-rates*, and *Codecs x Bit-rates* are the same as those already discussed in the previous section, but the main effect *Audio Content* and the interactions with *Audio Content* are of interest in this section. In Table 6 the main effect for *Audio Content* is significant, indicating that there was significant variation among the three classes of audio signals -- Mixed (65.87), Music (70.23), and Speech (66.29). Furthermore, the interactions *Codecs x Audio Content* and *Bit Rates x Audio Content* are also significant. These interactions are illustrated in Figs. 4a and 4b, respectively.

Table 6. Results of Audio Content ANOVA for Exp. 1-1 (Mono/Bit-rate)

Exp.1-1 - Cod x BR x AC (Lab rem.)	df		Sum of Squares		Mean Square		F-Ratio	Prob.
Codecs	1		94512		94512.0		288.30	0.0000
BitRates	2		341736		170868.0		914.50	0.0000
Votes	359		313682		873.8			
AC		2		8277		4138.5	4.84	0.0084
Votes w/n AC		357		305405		855.5		
Codecsx BitRates	2		25564		12782.0		97.46	0.0000
Codecs x Votes	359		120629		336.0			
Codecs x AC		2		3596		1798.0	5.48	0.0045
Codecs x Votes w/n AC		357		117033		327.8		
BitRates x Votes	718		140857		196.2			
BitRates x AC		4		7451		1862.8	9.97	0.0000
BitRates x Votes w/n AC		714		133406		186.8		
Codecs x BitRates x Votes	718		94225		131.2			
Codecs x BitRates x AC		4		581		145.3	1.11	0.3506
Codecs x BitRates x Votes w/n AC		714		93644		131.2		
Total	2159		1131205					

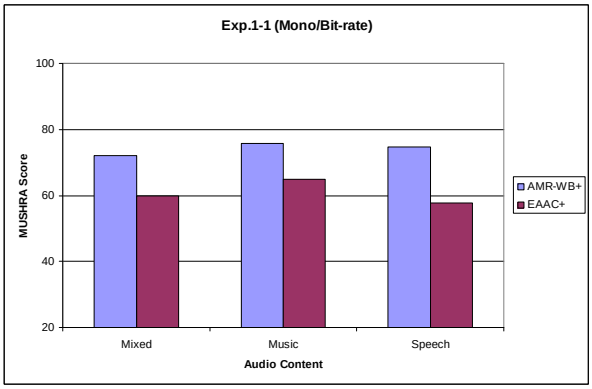


Fig.4a Interaction of Codecs x Audio Content.

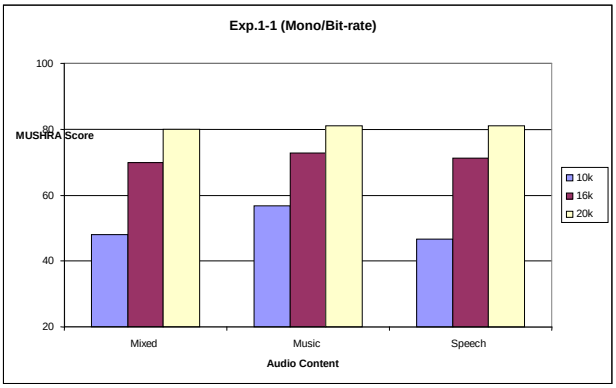


Fig.4b Interaction of Bit-rates x Audio Content.

Table 7 shows the results of the Audio Content ANOVA for the three codecs in Exp.1-1 operating at 10kbps, AMR-WB+, AMR-WB+(LC), and EAAC+. The ANOVA shows that both of the main effects, *Codecs* and *Audio Content* as well as the interaction *Codecs x Audio Content* are all significant. Figure 5a illustrates the significant main effects and Fig. 5b the significant interaction.

Table 7. Results of Codecs Audio Content ANOVA for Exp. 1-1 (Stereo/Bit-rate)

Exp.1-1 (Codecs at 10kbps)	df		Sum of Squares		Mean Square		F-ratio	Prob.
Codecs	2		112306.2		56153.08		219.01	0.0000
Votes	359		233598		650.69			
Audio Content		2		7509		3754.41	5.93	0.0029
Votes within AC		357		226089		633.30		
Codecs x Votes	718		193715.3		269.80			
Codecs x AC		4		10653		2663.24	10.39	0.0000
Codecs x Votes within AC		714		183062		256.39		
Total	1079		539619.5					

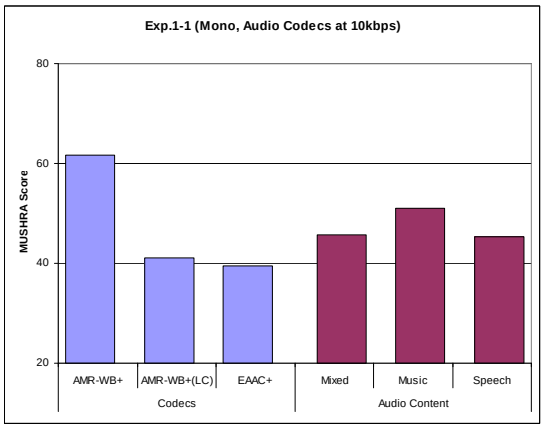


Fig.5a Main Effects for *Codecs* and *AC*

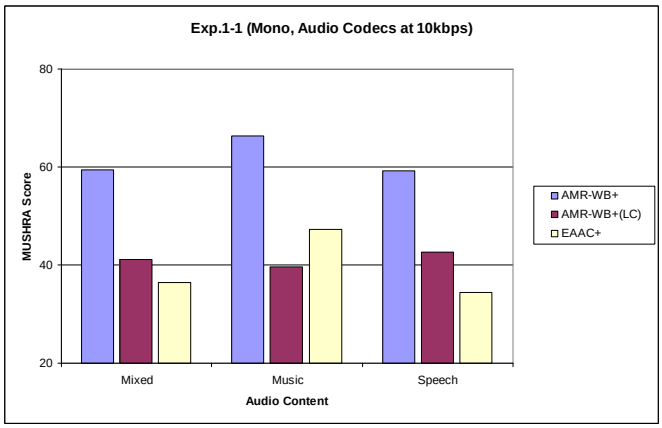


Fig.5b Interaction of *Codecs x Audio Content*.

Table 8 shows the ANOVA for Exp.1-2 for the effects of the factors *Codecs*, *Bit-rates*, and *Votes* with the fixed factor *Audio Content* nested within *Votes*. As in Table 6, the significance of the effects for *Codecs*, *Bit-rates*, and *Codecs x Bit-rates* are the same as those already discussed in the previous section. In Table 8 the main effect for *Audio Content* is not significant, indicating that there was no significant variation among the three classes of audio signals -- Mixed (50.96), Music (49.69), and Speech (51.65). However, all three of the interactions involving Audio Content were significant ($p<.05$) in the ANOVA. These significant interactions are illustrated in Fig. 6a (*Codecs x Audio Content*), Fig. 6b (*Bit-rates x Audio Content*), and Fig. 6c (*Codecs x Bit-rates x Audio Content*), respectively.

Table 8. Results of Audio Content ANOVA for Exp. 1-2 (Stereo/Bit-rate)

Exp.1-2 - Cod x BR x AC (Lab rem.)	df		Sum of Squares		Mean Square		F-Ratio	Prob.
Codecs	1		7		7.0		0.02	0.8876
BitRates	2		247441		123720.5		682.47	0.0000
Votes	359		266895		743.4			
AC		2		1384		691.8	0.93	0.3955
Votes w/n AC		357		265511		743.7		
Codecsx BitRates	2		1370		685.0		6.61	0.0014
Codecs x Votes	359		139466		388.5			
Codecs x AC		2		26805		13402.5	42.47	0.0000
Codecs x Votes w/n AC		357		112661		315.6		
BitRates x Votes	718		136353		189.9			
BitRates x AC		4		6918		1729.4	9.54	0.0000
BitRates x Votes w/n AC		714		129436		181.3		
Codecs x BitRates x Votes	718		80566		112.2			
Codecs x BitRates x AC		4		6531		1632.6	15.75	0.0000
Codecs x BitRates x Votes w/n AC		714		74036		103.7		
Total	2159		872098					

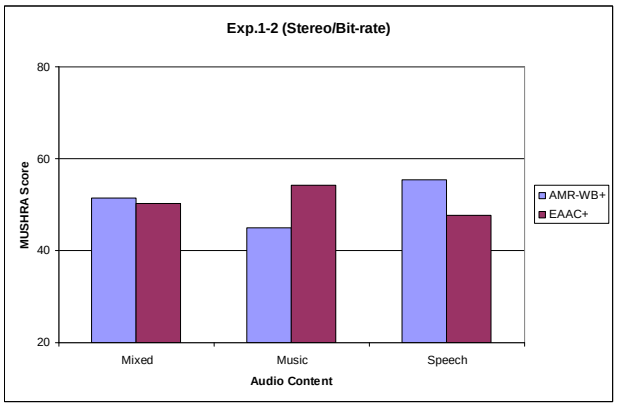


Fig.6a Interaction of Codecs x Audio Content.

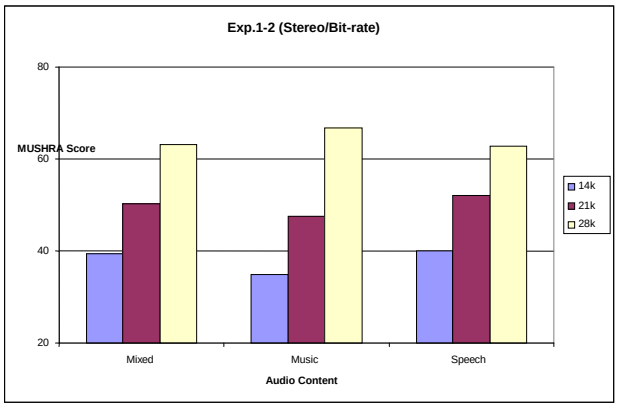


Fig.6b Interaction of Bit-rates x Audio Content.

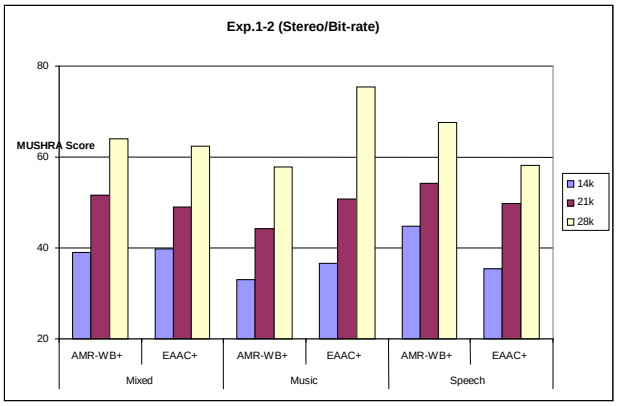


Fig.6c Interaction of Codecs x Bit-rates x Audio Content.

6. Conclusions

The conclusions listed below are based solely on the results and analyses derived from the Phase 1 MUSHRA tests and reported in this document. Except where explicitly stated, these conclusions apply only to the default AMR-WB+ codec since the low complexity version was only tested in one configuration -- at the 10kbps bit-rate in the Mono experiment.

- Codecs
  - Mono – the performance of Extended AMR-WB was significantly better than 3GPP Enhanced aacPlus across the three bit-rates
  - Stereo – performance of the two codecs were equivalent across the three bit-rates
  - Codecs x Bit-rate – 3GPP Enhanced aacPlus showed a greater improvement in performance than Extended AMR-WB with increases in Bit-rate in both Mono and Stereo
- Lab dependency
  - Mono - the difference between labs was significant but constant across conditions -- there were no significant interactions and therefore no Lab Dependency for Mono
  - Stereo - there were statistically significant interactions of *Codecs x Labs* and *Bit-rates x Labs* but the trends were consistent – there was a significant Lab Dependency effect

- Audio Content
  - Mono

- there were significant differences across classes of Audio Content
- Extended AMR-WB performed relatively better for Speech content, while 3GPP Enhanced aacPlus performed relatively better for Music content
- for the 10kbps bit-rate, the low complexity version of Extended AMR-WB performed better for Speech content than for Music, while 3GPP Enhanced aacPlus performed better for Music content than for Speech

- Stereo
  - there were no significant differences across classes of Audio Content
  - 3GPP Enhanced aacPlus showed a clear advantage over Extended AMR-WB with increases in bit-rate for Music content; Extended AMR-WB showed a clear advantage over 3GPP Enhanced aacPlus with increases in bit-rate for Speech content.

7. References

[1] S4-050188-PSS/MMS/MBMS Audio Codec Characterization Test Plan Version 0.5, Feb.2005.

[2] EBU Technical recommendation: MUSHRA-EBU Method for Subjective Listening Tests of Intermediate Audio Quality, Doc. B/AIM022, Oct.1999.