**Source: Dynastat[[1]](#footnote-2)1**

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

**Agenda item: 7**

# 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.

# 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 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



# 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



# 

### **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



# 

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

# 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, rLL = 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, Mean**Dyna** = 66.9, Mean**Eric** = 64.1) and somewhat larger for Exp.1-2 (Diff.=6.8, Mean**FTRD** = 48.5, Mean**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*.[[2]](#footnote-3) 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 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)



### 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)**



 

Fig.3a Interaction of Codecs x Bit Rates. Fig.3b Interaction of Codecs x Labs.



Fig.3c Interaction of Bit Rates x Labs.

# 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)**



 

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)**



 

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)**



 

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.

# 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.

# 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.

1. 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 [↑](#footnote-ref-2)
2. 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. [↑](#footnote-ref-3)