**Source: S. R. Quackenbush, Audio Research Labs**

**Title: Global Analysis Laboratory Report on 3GPP High-Rate Audio Codec Exercises**

**Status: Approved**

**Revision: March 5, 2004**

# Executive Summary

A series of three subjective listening tests were conducted as part of the 3GPP Audio codec exercise, as specified in document S4-030821, “PSS/MMS High-Rate Audio Selection Test and Processing Plan Version 2.2.0.” This documents reports the results of those tests.

The following table summarizes the performance of the codecs in the highest-rate of the Low-Rate tests for stereo signals on unimpaired channels (test A3 and A4, see S4-030824 [2] and S4-040173 [8]), and in each of the three High-Rate tests. In this table the two candidate codecs are AAC+ and CT. For each test, the codec with the best subjective score is highlighted in green, where “best” is in the statistical sense that the codec estimated mean score is better than that of the other codec at the 95% level of significance.

|  |  |  |  |
| --- | --- | --- | --- |
| Tests | Operating condition | AAC+ | CT |
| LR-A3 | 24 kbps, mono | 74.9 | 75.8 |
| LR-A4 | 24 kbps, stereo | 55.3 | 67.1 |
| 1 | 32 kbps, stereo | 75.8 | 84.9 |
| 2 | 48 kbps, stereo | 82.0 | 81.5 |
| 3-1 | 32 kbps, stereo, 1% FER | 66.2 | 72.9 |
| 3-2 | 32 kbps, stereo, 3% FER | 56.3 | 62.3 |

As the table shows, candidate CT appears to have consistently strong performance, having an estimated mean score at the 95% level of significance that is higher than that of candidate AAC+ in 4 of the 6 tests, and an estimated mean score that is not different from that of AAC+ in the remaining test.

The data support the following statements:

* Candidate CT had a mean score that was better than that of candidate AAC+ at the 95% level of significance in 4 of the 6 tests (LR-A4, 1, 2, 3-1, 3-2), and a mean score that is not different from that of AAC+ in the remaining tests (LR-A3, A2).

Table of Contents

[Executive Summary 1](#__RefHeading___Toc66239881)

[1 Introduction 3](#__RefHeading___Toc66239882)

[2 Overview of experiments 3](#__RefHeading___Toc66239883)

[3 Systems under test 3](#__RefHeading___Toc66239884)

[3.1 Candidate codecs 3](#__RefHeading___Toc66239885)

[3.2 Reference codecs 3](#__RefHeading___Toc66239886)

[3.3 Anchors and references 3](#__RefHeading___Toc66239887)

[4 Experimental design 4](#__RefHeading___Toc66239888)

[4.1 High-Rate Experiments 4](#__RefHeading___Toc66239889)

[4.2 Low-Rate Experiments applied to High-Rate Selection 5](#__RefHeading___Toc66239890)

[5 Test Material 5](#__RefHeading___Toc66239891)

[5.1 Signal categories 6](#__RefHeading___Toc66239892)

[5.2 Test Items 6](#__RefHeading___Toc66239893)

[5.3 Training Items 6](#__RefHeading___Toc66239894)

[6 Test sites 6](#__RefHeading___Toc66239895)

[7 Statistical analysis 6](#__RefHeading___Toc66239896)

[7.1 Overview 7](#__RefHeading___Toc66239897)

[7.2 Statistical Model Based on the Experimental Design 7](#__RefHeading___Toc66239898)

[7.3 Pivot Table and ANOVA Analysis 8](#__RefHeading___Toc66239899)

[7.4 Post-Processing of Listener Data 9](#__RefHeading___Toc66239900)

[7.5 Analysis Process 9](#__RefHeading___Toc66239901)

[8 Test Results 10](#__RefHeading___Toc66239902)

[8.1 Test 1 10](#__RefHeading___Toc66239903)

[8.2 Test 2 16](#__RefHeading___Toc66239904)

[8.3 Test 3 22](#__RefHeading___Toc66239905)

[9 Application of Selection Rules 29](#__RefHeading___Toc66239906)

[9.1 Selection Rule 1 29](#__RefHeading___Toc66239907)

[9.2 Selection Rule 2 29](#__RefHeading___Toc66239908)

[9.3 Selection Rule 3 29](#__RefHeading___Toc66239909)

[Reference Documents 31](#__RefHeading___Toc66239910)

# Introduction

The European Telecommunications Standards Institute (ETSI) has conducting a series of subjective listening tests as part of the 3GPP Audio codec exercise. 3GPP desires to use the tests to evaluate candidate codecs for their needs, as set forth in documents S4-030821, “PSS/MMS High-Rate Audio Selection Test and Processing Plan Version 2.2.0” [1] and S4-030824, “AMR-WB+ and PSS/MMS Low-Rate Audio Selection Test and Processing Plan Version 2.2” [2]. This documents reports the results of those tests.

# Overview of experiments

The High-Rate tests were comprised of three experiments defined in [1]. The Selection Rules (Section 9) uses the results of two additional experiments defined in [2].

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Exp.** | **Operational mode** | **#Codecs in test** | **# reference codecs** | **#Anchors in test** | **#References** | **#items** | **Total** |
| 1 | 32 kbps, stereo | 2(use case B encoder) | 2, incl. RealAudio @ 32 kbit/s stereo | 2 | 1 | 12 | 84 |
| 2 | 48 kbps, stereo | 2(use case B encoder) | 2, incl. RealAudio @ 48 kbit/s stereo | 2 | 1 | 12 | 84 |
| 3 | 32 kbps, stereo, 1% and 3% random frame loss | 4 (2 candidates at 2 frame loss rates each) | 2 (AAC-LC at 2 frame loss rates) | 2 | 1 | 12 | 108 |

# Systems under test

## Candidate codecs

The candidate codec participating in the PSS/MMS high-rate audio selection tests are listed in the following table.

|  |  |  |
| --- | --- | --- |
| **#** | **Codec name** | **Providing Organization(s)** |
| 1 | AAC+ | Coding Technologies/ NEC |
| 2 | CT | Coding Technologies |

## Reference codecs

The reference codecs are listed in the following table.

|  |  |  |
| --- | --- | --- |
| **#** | **Codec name** | **Providing Organization(s)** |
| 3 | AAC | Fraunhofer |
| 4 | RealAudio | RealNetworks |

## Anchors and references

Besides the items encoded with the candidate and reference codecs, anchor and reference items were included in the tests. In the experiments, two anchors will be used with lowpass filtered original signal.

Also included is the uncoded original signal, once as open and once as hidden reference.

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Type** | **Specification** | **Channel type** |
| 1 | Anchor | 3.5 kHz Lowpass | Mono and Stereo |
| 2 | Anchor | 7.0 kHz Lowpass | Mono and Stereo |
| 6 | Hidden Reference | Original signal | Mono and Stereo |
| 7 | Open Reference | Original signal | Mono and Stereo |

# Experimental design

The following tables show the parameters, candidate codes, reference codecs and anchors and references for each experiment. The row labels in the first column (headed “Parameter”) are explained as follows:

* The row labeled “Experiment” indicates the experiment. Each experiment is specified in a separate table.
* The row labeled “Bit Rate” indicates the bitrate for the experiment.
* The row labeled “Signal” indicates the number of distinct channels in the test material (i.e. mono or stereo).
* The row labeled “Candidate codecs” lists each candidate codec tested in the experiment in sub-divisions of that row. All Candidate codecs process 48 kHz sampling rate test material and code at bit rate indicated for each experiment unless explicitly indicated otherwise.
* The row labeled “Reference codecs” lists each reference codec tested in the experiment in sub-divisions of that row. All Reference codecs process 48 kHz sampling rate test material and code at bit rate indicated for each experiment unless explicitly indicated otherwise (e.g. RealAudio in experiment 1).
* The row labeled “Anchors and references” lists each anchor and reference condition tested in the experiment in sub-divisions of the main row.

## High-Rate Experiments

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Additional Constraints** |
| Experiment | 1 |  |
| Bit Rate | 32 kbps |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+ |  |
| CT |  |
| Reference codecs | AAC |  |
| RealAudio | 22.05 kHz sampling rate |
| Anchors and references | Open Reference |  |
| Hidden Reference |  |
| 7.0 kHz Lowpass |  |
| 3.5 kHz Lowpass |  |

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Additional Constraints** |
| Experiment | 2 |  |
| Bit Rate | 48 kbps |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+ |  |
| CT |  |
| Reference codecs | AAC |  |
| RealAudio | 44.1 kHz sampling rate |
| Anchors and references | Open Reference |  |
| Hidden Reference |  |
| 7.0 kHz Lowpass |  |
| 3.5 kHz Lowpass |  |

Experiment 3 simulated errored channels using two conditions, 1 percent frame error rate (FER) and 3 percent FER. The application of the two error conditions doubled the number of systems under test. Note, however, that the RealAudio reference codec was not present in this experiment.

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Additional Constraints** |
| Experiment | 3 |  |
| Bit Rate | 32 kbps |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+ FER 1% |  |
| AAC+ FER 3% |  |
| CT FER 1% |  |
| CT FER 3% |  |
| Reference codecs | AAC FER 1% |  |
| AAC FER 3% |  |
| Anchors and references | Open Reference |  |
| Hidden Reference |  |
| 7.0 kHz Lowpass |  |
| 3.5 kHz Lowpass |  |

## Low-Rate Experiments applied to High-Rate Selection

For more details on these experiments see [2].

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Additional Constraints** |
| Experiment | A3a and A3b |  |
| Bit Rate | 24 kbps |  |
| Signal | Mono |  |
| Candidate codecs | AAC+ |  |
| AMR-WB+ |  |
| CT |  |
| Reference codecs | AAC |  |
| AMR-WB | 23.85 kbps, 16 kHz sampling rate |
| Anchors and references | Open Reference |  |
| Hidden Reference |  |
| 7.0 kHz Lowpass |  |
| 3.5 kHz Lowpass |  |

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Additional Constraints** |
| Experiment | A4a and A4b |  |
| Bit Rate | 24 kbps |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+ |  |
| AMR-WB+ |  |
| CT |  |
| Reference codecs | AAC |  |
| AMR-WB | 23.85 kbps, 16 kHz sampling rate |
| Anchors and references | Open Reference |  |
| Hidden Reference |  |
| 7.0 kHz Lowpass | 6 dB attenuated side channel |
| 7.0 kHz Lowpass | 2 dB attenuated side channel |
| 3.5 kHz Lowpass | 12 dB attenuated side channel |

# Test Material

## Signal categories

The test material was selected so as to be representative of the following signal categories:

* Classic, with and/or without vocals
* Pop, with and/or without vocals
* Single instruments
* Mixed speech and music
* Speech with and/or without background noise
* a capella vocals, solo and/or choir

## Test Items

A single set of twelve test items were used for the three experiments. They are:

c\_01\_org.wav

c\_02\_org.wav

p\_01\_org.wav

p\_02\_org.wav

si\_01\_org.wav

si\_02\_org.wav

sm\_01\_org.wav

sm\_02\_org.wav

sp\_01\_org.wav

sp\_02\_org.wav

sp\_03\_org.wav

v\_01\_org.wav

Original material was in stereo, and for mono experiments it was downmixed.

## Training Items

A single set of four training items are used for the three tests. They are:

c\_09\_org.wav

p\_09\_org.wav

si\_09\_org.wav

sp\_09\_org.wav

# Test sites

The experiments for each candidate codec are run by two listening laboratories in parallel, as shown in Table 6-1.

Table 6‑1: Allocation of sub-experiments to the Listening Laboratories

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Exp. | Lab1 | Lab2 | Lab3 | Lab4 | Lab5 | Lab6 | Total |
| LL ID | TS | NT | FT | DY | NK | ER | Per Exp. |
| 1 | X |  |  | X |  |  | 2 |
| 2 |  | x |  |  | x |  | 2 |
| 3 |  |  | x |  |  | x | 2 |
| Totals: | 1 | 1 | 1 | 1 | 1 | 1 | 6 |

(Legend: T-Systems (TS), NTT-AT (NT), France Telecom R&D (FT), Dynastat (DY), Nokia (NK),   
Ericsson (ER)

# Statistical analysis

## Overview

### Standard Pivot Table Analysis

The Pivot Table statistical analysis followed the standard MUSHRA procedure [3].

The calculation of the averages of the scores of all listeners remaining after post-screening will result in the Mean Subjective Scores (MSS).

The first step of the analysis of the results is the calculation of the mean score , for each of the presentations:



where:

is the score of observer i for a given test condition j and sequence k

*N* is the number of observers

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



where:



and the standard deviation is given by: 

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

Similarly, a standard deviation is calculated for each test condition. It is noted however that this standard deviation may be influenced more by differences associated with the test sequences than by differences associated with the listeners participating in the assessment.

## Statistical Model Based on the Experimental Design

The basic model of a score can be thought of as the sum of “effects”. A particular score may depend on which codec was involved, which audio selection is being played, which laboratory is conducting the test, and which subject is listening.

We anticipate, *a priori*, that there may also be an interaction between the audio selection and the codec under test. In other words, some codecs may perform better with some types of audio selections than with others. Further, we anticipate, *a priori*, that there may also be an interaction between the codecs under test and the testing laboratory. The proposed analysis evaluates whether these interactions exist and compensates for them, if necessary.

Further, in statistical terminology, subjects are “nested” within laboratories. In other words, subject 1 in laboratory A is a different person, with different characteristics, from subject 1 in laboratory B.

Using a simple notation, the proposed basic model for the high-rate experiments as described above is

Score = Codec (c = 1, …, 7 or 9)

+ Signal Category (SigCat = 1, … 6)

+ Signal (Signal = 1, …, 12)

+ Codec by Signal Category interaction

(Codec:SigCat, Codec = 1, …, 7 or 9, SigCat = 1, …, 6)

+ Laboratory (Site = 1, …, 2)

+ Codec by Laboratory interaction (Codec:Site, Codec = 1, …, 7 or 9, Site = 1, …, 2)

+ Subjects (s = 1, …, 15 for each Site)

+ Experimental error

In other words, the score is the sum of a number of factors plus random “error.” Just the codec main effects, and possibly the codec by signal category interaction are of real interest. The main effects are analogues of the Pivot Table averages. The interaction term for, say, the codec by signal category interaction takes into account that a response might not be predictable simply by adding an effect for the codec and an effect for the signal category. Some codecs may be “winners” for some signal category, while other codecs may be “winners” for other signal categories. The statistical significance and the size of these effects will be a measure of how important the interaction terms are

There will be one instance of this model for each of the 3 high-rate experiments.

The experimental design is balanced, in that there are equal numbers of each factor level involved with each codec, with the exception that the signal categories are not equally represented. This balance has the advantage that the mean score for each codec is an appropriate statistic for estimating the quality of that codec, assuming that the signal categories are close to balanced. As discussed below, it is the estimates of the standard deviations (or equivalently, the widths of the confidence intervals) that are different depending on the method of analysis. It would be best to use the analysis method that yields the narrowest confidence intervals, thereby giving the most information for the money spent.

Further, as mentioned in the Analysis Process section below, some Subject-Signal judgments of the codecs will be eliminated because they appear to be inconsistent with *a priori* expectations. To the extent that this happens, the analysis of variance will have to compensate for this imbalance.

## Pivot Table and ANOVA Analysis

Data from experiments such of these have been analyzed in the past using the Pivot Table facilities of MS Excel spreadsheets. For simple experiments, this is probably adequate. However, the experiments being analyzed in these tests are far from simple. The pivot table is used to calculate for each codec a grand average (across all signals, subjects, etc.) and the standard deviation of that average. From these, confidence intervals can be constructed, and differences between codecs can be evaluated.

The problem from a statistical viewpoint with this analysis for the experiments described here is that the standard deviations are inflated by the variability of the other factors. This results in a test with less statistical resolving power. In other words, for a given confidence interval width, the Pivot Table method of analysis requires more listeners than the analysis method described here, or, for a given number of listeners, the proposed analysis of variance method yields narrower confidence intervals than the Pivot Table method. The reason for this is that, for example, although each codec is rated for each signal, and therefore the signal differences cancel out when comparing averages, the difference between signals will make the numbers gathered into that average more variable than they would be if the average signal effects were subtracted out first.

The statistical technique called Analysis of Variance or ANOVA will perform the appropriate analysis, better estimating the standard deviations and confidence intervals for the differences between codecs. A detailed description of ANOVA is beyond the scope of this document, but references are given in Section 7.5

## Post-Processing of Listener Data

The MUSHRA test methodology provides very limited ability to assess the reliability of individual listeners. In this analysis, listener reliability was assessed by observing the extent to which the listener scored the hidden reference at 100 and gave monotonically decreasing scores to each of the hidden reference, the 7.5 kHz lowpass anchor and the 3.6 kHz lowpass anchor. An interval for modest listener error was allowed in applying this rule, e.g. that the hidden reference must be scored higher than 85 rather than exactly 100. Similarly, scores may depart from strict monotonicity by 10 points and still be allowed. These values (85 and 10) were chosen to allow for more listener error than in the low rate experiments because the differences in quality of the high rate signals appeared to be harder to judge than with the low rate signals.

## Analysis Process

The analysis will proceed through the following steps

1. The MS Excel data templates are prepared in the approved format.
2. The data arrives from the testing laboratories in the MS Excel data template.
3. The data from the both labs is compiled into a single workbook for each experiment.
4. A Visual Basic program is used to unstack the data so that each row will have only one listener response.
5. The condition labels are replaced by the correct, unrandomized codec names.
6. A consistency check is performed. Listener-signal combinations are eliminated (given a Weight of 0) if
   * the hidden reference does not receive a rating of at least 85 or
   * the lp3500 anchor rating is not more than 10 units greater than the lp7000 anchor rating.
7. A Pivot Table analysis is performed to obtain simple, benchmark results, from which appropriate presentation charts are created. As described above, the more complex ANOVA analysis should produce codec means which are very close to the pivot table means, differing only in the effect of any missing or eliminated data. The main difference in results will be that the ANOVA confidence intervals will be narrower than the Pivot Table confidence intervals.
8. The data is exported to a text file and entered into “R” [4], a gnu version of the statistical analysis system called “S” [5]. A script is used to fit the model. In particular, the function aov() [6] is used to fit a linear model (the ANOVA model above) to the data. The fitted codec effects and interactions, estimated standard errors of the effects, and the estimated standard error of the residuals are used to create the appropriate confidence intervals. The output from R is captured in a text file.
9. The Visual Basic programs used to compile and screen the data, Excel workbook with all received data and the Pivot Table analysis, the R analysis script, and the text file of R output are all available as part of this report.

# Test Results

In this section the candidate codecs are named only in the initial table showing test parameters. In all subsequent data analysis they are referred to using the labels Codec1 and Codec2 such that their identity is concealed.

## Test 1

### Test parameters and systems under test

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Symbol** |
| Experiment | 1 |  |
| Bit Rate | 32 kbps |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+ | Codec1 |
| CT | Codec2 |
| Reference codecs | AAC | AAC |
| RealAudio@32 kbit/s stereo | RN |
| Anchors and references | Open Reference |  |
| Hidden Reference | hidref |
| 7.0 kHz Lowpass | LP7.0 |
| 3.5 kHz Lowpass | LP3.5 |

### Pivot Table Results

The following chart shows the overall relative performance of the codecs in this experiment. The means and 95% confidence intervals shown are from the standard Pivot Table analysis in which the summary statistics are computed over all signals listeners, and laboratories.



Each of the candidate codecs out-performs both of the reference codecs. The following table shows the numerical values plotted in the chart above.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | AAC | hidref | lp3500 | lp7000 | RN |
| Average | 75.8 | 84.9 | 38.7 | 99.6 | 26.7 | 53.6 | 48.0 |
| Lower Bound | 73.5 | 83.2 | 36.0 | 99.4 | 24.4 | 50.9 | 45.2 |
| Upper Bound | 78.1 | 86.5 | 41.5 | 99.8 | 29.1 | 56.2 | 50.8 |

The following 2 charts show the performance of each of the candidate codecs for each of the test signals.





The following table presents the data used to create the previous charts.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Codec 1 | | | Codec 2 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| c\_1 | 81.8 | 66.6 | 74.2 | 90.7 | 82.2 | 86.4 |
| c\_2 | 92.1 | 83.2 | 87.6 | 91.0 | 81.4 | 86.2 |
| p\_1 | 81.3 | 68.3 | 74.8 | 92.0 | 78.8 | 85.4 |
| p\_2 | 84.0 | 68.0 | 76.0 | 95.1 | 88.4 | 91.7 |
| si\_1 | 87.6 | 74.7 | 81.2 | 90.3 | 76.4 | 83.3 |
| si\_2 | 89.7 | 71.2 | 80.5 | 89.1 | 78.1 | 83.6 |
| sm\_1 | 85.9 | 72.7 | 79.3 | 93.6 | 86.4 | 90.0 |
| sm\_2 | 76.8 | 64.3 | 70.6 | 87.0 | 74.5 | 80.7 |
| sp\_1 | 77.1 | 61.3 | 69.2 | 80.7 | 66.2 | 73.5 |
| sp\_2 | 90.1 | 81.6 | 85.9 | 89.8 | 78.2 | 84.0 |
| sp\_3 | 63.5 | 42.1 | 52.8 | 93.8 | 86.0 | 89.9 |
| v\_1 | 86.1 | 70.4 | 78.2 | 89.6 | 78.4 | 84.0 |

### Analysis of Variance Results

The data were analyzed using Analysis of Variance techniques. The following are the overall basic results from the Analysis of Variance:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Codec | 8 | 1326938 | 165867 | 856.4 | < 2.2e-16 \*\*\* |
| SigCat | 5 | 15238 | 3048 | 15.7 | 2.50E-15 \*\*\* |
| Signal | 6 | 40742 | 6790 | 35.1 | < 2.2e-16 \*\*\* |
| Site | 1 | 109184 | 109184 | 563.7 | < 2.2e-16 \*\*\* |
| Subject | 28 | 182687 | 6525 | 33.7 | < 2.2e-16 \*\*\* |
| Codec:Signal | 40 | 36003 | 900 | 4.6 | < 2.2e-16 \*\*\* |
| Codec:Site | 8 | 20330 | 2541 | 13.1 | < 2.2e-16 \*\*\* |
| Residuals | 3125 | 605265 | 194 |  |  |

Signif. codes: 0 < \*\*\* < 0.001 < \*\* < 0.01 < \* < 0.05 < • < 0.1 < ‘ ‘ < 1

All components of the model are highly statistically significant at greater than the 99.9% level. This means that each of the aspects of the experimental design was important and rightfully included in the model, so that the effect of that component can be compensated for when analyzing the variable of interest, the difference between the codecs. However, it should be kept in mind that this experiment resulted in much data being collected, and small differences can be statistically significant, while their practical effect is minimal.

The following are the main effects (the estimated mean of each level of each variable) as determined by this analysis.

Codec main effect

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | AAC | RN | hidref | lp3500 | lp7000 |
| mean | 75.8 | 84.9 | 38.7 | 48.0 | 99.6 | 26.7 | 53.6 |
| N | 354 | 354 | 354 | 354 | 354 | 354 | 354 |
| Lower Bound | 74.4 | 83.5 | 37.3 | 46.6 | 98.2 | 25.3 | 52.2 |
| Upper Bound | 77.2 | 86.3 | 40.1 | 49.4 | 101.0 | 28.1 | 55.0 |

As can be seen by comparing this table with the Pivot Table analysis means above, the two analyses give almost identical results. As mentioned, the difference between the analyses is in the width of the confidence intervals.

Signal Category main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | c | p | si | sm | sp | v |
| mean | 61.1 | 58.9 | 64.7 | 60.8 | 60.1 | 61.5 |
| N | 413 | 413 | 406 | 413 | 623 | 210 |

Although this variable is highly statistically significant, the signal categories have means that do not differ too much. The practical differences may not be too great. The statistical significance here means that the largest mean is definitely statistically significantly different from the smallest, but other differences would require a more in-depth analysis.

Codec by Signal Category (Codec:SigCat) interaction effect

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Codec | SigCat | | | | | | |
|  |  | c | p | si | sm | sp | v |
| Codec1 | mean | 81.1 | 75.4 | 80.8 | 74.9 | 69.1 | 78.2 |
| rep | N | 59 | 59 | 58 | 59 | 89 | 30 |
| Codec2 | mean | 86.3 | 88.5 | 83.5 | 85.3 | 82.5 | 84.0 |
| rep | N | 59 | 59 | 58 | 59 | 89 | 30 |
| AAC | mean | 38.9 | 36.0 | 47.5 | 39.1 | 32.3 | 45.3 |
| rep | N | 59 | 59 | 58 | 59 | 89 | 30 |
| RN | mean | 43.6 | 43.5 | 47.8 | 50.1 | 54.9 | 40.8 |
| rep | N | 59 | 59 | 58 | 59 | 89 | 30 |
| hidref | mean | 99.2 | 99.6 | 99.5 | 100.0 | 99.6 | 99.8 |
| rep | N | 59 | 59 | 58 | 59 | 89 | 30 |
| lp3500 | mean | 27.5 | 23.9 | 31.1 | 24.7 | 26.3 | 27.7 |
| rep | N | 59 | 59 | 58 | 59 | 89 | 30 |
| lp7000 | mean | 51.0 | 45.2 | 62.5 | 51.2 | 56.2 | 54.9 |
|  | N | 59 | 59 | 58 | 59 | 89 | 30 |

As can be seen in the above table, some codecs perform relatively better in some signal categories, while other codecs perform better in other signal categories. This is the meaning of “interaction.” The set of codec by signal category interactions above are statistically significant. Without presenting all the confidence intervals, the width of the 95% confidence intervals for the sp category is ±2.8, while the width of the 95% confidence intervals for the v category is ±4.8, and the width of the 95% confidence intervals for the other categories is ±3.4.

Signal main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | c\_1 | c\_2 | p\_1 | p\_2 | si\_1 | si\_2 |
| mean | 57.5 | 64.5 | 58.6 | 63.6 | 57.0 | 65.1 |
| N | 203 | 210 | 210 | 203 | 203 | 203 |
|  | sm\_1 | sm\_2 | sp\_1 | sp\_2 | sp\_3 | v\_1 |
| mean | 63.9 | 58.3 | 59.5 | 66.3 | 57.5 | 61.04 |
| N | 203 | 210 | 210 | 203 | 210 | 210 |

The signal main effects are shown here for completeness. The differences are statistically significant, but since the each signal is a unique item, it is not clear what use can be made of these individual means.

Site main effect

|  |  |  |
| --- | --- | --- |
|  | DY | T-Sys |
| mean | 74.6 | 47.7 |
| N | 1232 | 1246 |

The sites are statistically significantly different. Again, it is not clear what use can be made of these individual means.

Subject main effect

The subjects are statistically significantly different. The details of subject results can be found in the accompanying spreadsheets..

### Sources of variability

There is definitely a statistically significant and practically significant interaction between codecs and signals. That is, some codecs worked better for some signals than for others. These interactions can best be reviewed by studying the three charts above where, for each codec under test, the quality ratings are shown for each signal.

There is also definitely a statistically significant codec by lab interaction. In other words, some codecs performed relatively better in some testing labs than in others.

### Post-screening of data

Of the 360 sets of 7 judgments (one for each codec, reference codec, and anchor) in this experiment, 6 were eliminated by the post-screening procedure. The results of the screening procedure are coded by the Weight variable, where passing judgments received a 1 and eliminated judgments received a 0. In the pivot table, this variable can be manipulated to show the Pivot Table results with all the data. The means do not change much in a practical sense.

## Test 2

### Test parameters and systems under test

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Symbol** |
| Experiment | 2 |  |
| Bit Rate | 48 kbps |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+ | Codec1 |
| CT | Codec2 |
| Reference codecs | AAC | AAC |
| RealAudio@48 kbit/s stereo | RN |
| Anchors and references | Open Reference |  |
| Hidden Reference | hidref |
| 7.0 kHz Lowpass | LP7.0 |
| 3.5 kHz Lowpass | LP3.5 |

### Pivot Table Results

The following chart shows the overall relative performance of the codecs in this experiment. The means and 95% confidence intervals shown are from the standard Pivot Table analysis in which the summary statistics are computed over all signals listeners, and laboratories.



Each of the candidate codecs out-performs both of the reference codecs.

The following table shows the numerical values plotted in the chart above.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | AAC | hidref | lp3500 | lp7000 | RN |
| Average | 82.0 | 81.5 | 60.5 | 98.7 | 27.1 | 45.4 | 64.1 |
| Lower Bound | 80.0 | 79.5 | 57.7 | 98.3 | 25.2 | 43.2 | 61.6 |
| Upper Bound | 84.1 | 83.5 | 63.3 | 99.0 | 29.0 | 47.6 | 66.7 |

The following 2 charts show the performance of each of the candidate codecs for each of the test signals.





The following table presents the data used to create the previous charts.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Codec 1 | | | Codec 2 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| c\_1 | 86.9 | 74.9 | 80.9 | 84.8 | 73.0 | 78.9 |
| c\_2 | 96.8 | 82.0 | 89.4 | 96.8 | 91.2 | 94.0 |
| p\_1 | 88.6 | 75.3 | 82.0 | 88.5 | 79.7 | 84.1 |
| p\_2 | 89.8 | 77.5 | 83.7 | 93.6 | 84.0 | 88.8 |
| si\_1 | 92.2 | 82.0 | 87.1 | 92.4 | 82.5 | 87.4 |
| si\_2 | 96.1 | 87.2 | 91.6 | 93.0 | 85.4 | 89.2 |
| sm\_1 | 95.0 | 85.5 | 90.3 | 91.0 | 80.6 | 85.8 |
| sm\_2 | 85.1 | 69.1 | 77.1 | 86.5 | 69.9 | 78.2 |
| sp\_1 | 74.9 | 57.0 | 65.9 | 72.3 | 55.6 | 64.0 |
| sp\_2 | 94.8 | 82.7 | 88.7 | 89.8 | 76.7 | 83.3 |
| sp\_3 | 72.4 | 57.7 | 65.0 | 68.9 | 52.1 | 60.5 |
| v\_1 | 88.9 | 77.0 | 83.0 | 90.1 | 78.6 | 84.3 |

### Analysis of Variance Results

The data were analyzed using Analysis of Variance techniques. The following are the overall basic results from the Analysis of Variance:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Codec | 6 | 1148537 | 191423 | 785.6 | < 2.2e-16 \*\*\* |
| SigCat | 5 | 13303 | 2661 | 10.9 | 2.13e-10 \*\*\* |
| Signal | 6 | 28346 | 4724 | 19.4 | < 2.2e-16 \*\*\* |
| Site | 1 | 1 | 1 | 0.0 | 0.96 |
| Subject | 28 | 216419 | 7729 | 31.7 | < 2.2e-16 \*\*\* |
| Codec:Signal | 30 | 62531 | 2084 | 8.6 | < 2.2e-16 \*\*\* |
| Codec:Site | 6 | 4127 | 688 | 2.8 | 0.01 \*\* |
| Residuals | 2192 | 534086 | 244 |  |  |

Signif. codes: 0 < \*\*\* < 0.001 < \*\* < 0.01 < \* < 0.05 < • < 0.1 < ‘ ‘ < 1

All components of the model are highly statistically significant at greater than the 99.9% level except Site, which is not significant, and the Codec by Site interaction, which is statistically significant at the 99% level. This means that each of the aspects of the experimental design, except possibly Site, was important and rightfully included in the model, so that the effect of that component can be compensated for when analyzing the variable of interest, the difference between the codecs. However, it should be kept in mind that this experiment resulted in much data being collected, and small differences can be statistically significant, while their practical effect is minimal.

The following are the main effects (the estimated mean of each level of each variable) as determined by this analysis.

Codec main effect

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | AAC | RN | hidref | lp3500 | lp7000 |
| mean | 82.0 | 81.5 | 60.5 | 64.1 | 98.7 | 27.1 | 45.4 |
| N | 325 | 325 | 325 | 325 | 325 | 325 | 325 |
| Lower Bound | 80.3 | 79.8 | 58.8 | 62.4 | 97.0 | 25.4 | 43.7 |
| Upper Bound | 83.7 | 83.2 | 62.2 | 65.8 | 100.4 | 28.8 | 47.1 |

As can be seen by comparing this table with the Pivot Table analysis means above, the two analyses give almost identical results. As mentioned, the difference between the analyses is in the width of the confidence intervals.

Signal Category main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | c | p | si | sm | sp | v |
| mean | 67.3 | 66.8 | 67.7 | 66.7 | 61.5 | 66.2 |
| N | 364 | 371 | 385 | 378 | 581 | 196 |

Although this variable is highly statistically significant, the signal categories have means that do not differ too much. The practical differences may not be too great. The statistical significance here means that the largest mean is definitely statistically significantly different from the smallest, but other differences would require a more in-depth analysis.

Codec by Signal Category (Codec:SigCat) interaction effect

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Codec | SigCat | | | | | | |
|  |  | c | p | si | sm | sp | v |
| Codec1 | mean | 85.3 | 82.8 | 89.4 | 83.9 | 73.1 | 83.0 |
|  | N | 52 | 53 | 55 | 54 | 83 | 28 |
| Codec2 | mean | 86.8 | 86.5 | 88.3 | 82.2 | 69.1 | 84.3 |
|  | N | 52 | 53 | 55 | 54 | 83 | 28 |
| AAC | mean | 78.2 | 60.5 | 52.5 | 62.8 | 49.3 | 71.5 |
|  | N | 52 | 53 | 55 | 54 | 83 | 28 |
| RN | mean | 56.3 | 67.6 | 63.0 | 69.0 | 67.5 | 54.6 |
|  | N | 52 | 53 | 55 | 54 | 83 | 28 |
| hidref | mean | 98.0 | 98.8 | 98.4 | 98.5 | 99.4 | 98.5 |
|  | N | 52 | 53 | 55 | 54 | 83 | 28 |
| lp3500 | mean | 25.3 | 27.4 | 29.9 | 27.2 | 26.0 | 27.9 |
|  | N | 52 | 53 | 55 | 54 | 83 | 28 |
| lp7000 | mean | 41.4 | 44.1 | 51.9 | 43.3 | 46.4 | 43.6 |
|  | N | 52 | 53 | 55 | 54 | 83 | 28 |

As can be seen in the above table, some codecs perform relatively better in some signal categories, while other codecs perform better in other signal categories. This is the meaning of “interaction.” The set of codec by signal category interactions above are statistically significant. Without presenting all the confidence intervals, the width of the 95% confidence intervals for the sp category is ±3.4, while the width of the 95% confidence intervals for the v category is ±5.8, and the width of the 95% confidence intervals for the other categories is ±4.2.

Signal main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | c\_1 | c\_2 | p\_1 | p\_2 | si\_1 | si\_2 |
| mean | 64.7 | 66.5 | 64.1 | 67.1 | 61.2 | 69.9 |
| N | 175 | 189 | 182 | 189 | 189 | 196 |
|  | sm\_1 | sm\_2 | sp\_1 | sp\_2 | sp\_3 | v\_1 |
| mean | 68.5 | 62.6 | 63.8 | 73.0 | 60.2 | 65.61 |
| N | 196 | 182 | 203 | 189 | 189 | 196 |

The signal main effects are shown here for completeness. The differences are statistically significant, but since the each signal is a unique item, it is not clear what use can be made of these individual means.

Site main effect

|  |  |  |
| --- | --- | --- |
|  | Nokia | NTT-AT |
| mean | 65.63 | 65.6 |
| N | 1183 | 1092 |

The sites are not statistically significantly different, although the interaction between sites and codecs is statistically significant at the 99% level.

Subject main effect

The subjects are statistically significantly different. The details of subject results can be found in the accompanying spreadsheets..

### Sources of variability

There is definitely a statistically significant and practically significant interaction between codecs and signals. That is, some codecs worked better for some signals than for others. These interactions can best be reviewed by studying the three charts above where, for each codec under test, the quality ratings are shown for each signal.

There is also definitely a statistically significant codec by lab interaction. In other words, some codecs performed relatively better in some testing labs than in others.

### Post-screening of data

Of the 360 sets of 7 judgments (one for each codec, reference codec, and anchor) in this experiment, 35 were eliminated by the post-screening procedure. The results of the screening procedure are coded by the Weight variable, where passing judgments received a 1 and eliminated judgments received a 0. In the pivot table, this variable can be manipulated to show the Pivot Table results with all the data. The means change less than 1 unit, which is not much in a practical sense.

## Test 3

### Test parameters and systems under test

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Symbol** |
| Experiment | 3 |  |
| Bit Rate | 32 kbps, 1% and 3% random frame loss |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+, 1% random frame loss | Codec1\_FER1 |
| AAC+, 3% random frame loss | Codec1\_FER3 |
| CT, 1% random frame loss | Codec2\_FER1 |
| CT, 3% random frame loss | Codec2\_FER3 |
| Reference codecs | AAC, 1% random frame loss | AAC\_FER1 |
| AAC, 3% random frame loss | AAC\_FER3 |
| Anchors and references | Open Reference |  |
| Hidden Reference | hidref |
| 7.0 kHz Lowpass | LP7.0 |
| 3.5 kHz Lowpass | LP3.5 |

### Pivot Table Results

The following chart shows the overall relative performance of the codecs in this experiment. The means and 95% confidence intervals shown are from the standard Pivot Table analysis in which the summary statistics are computed over all signals listeners, and laboratories.



Each of the candidate codecs out-performs both of the reference codecs. The following table shows the numerical values plotted in the chart above.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1\_FER1 | Codec1\_FER3 | Codec2\_FER1 | Codec2\_FER3 | AAC\_ FER1 | AAC\_ FER3 | hidref | lp3500 | lp7000 |
| Average | 66.2 | 56.3 | 72.9 | 62.3 | 38.7 | 33.7 | 99.8 | 31.7 | 57.2 |
| Lower Bound | 64.1 | 54.1 | 71.0 | 60.0 | 36.8 | 32.1 | 99.6 | 30.1 | 55.3 |
| Upper Bound | 68.2 | 58.5 | 74.8 | 64.6 | 40.5 | 35.4 | 100.0 | 33.4 | 59.1 |

The following 4 charts show the performance of each of the candidate codecs for each of the test signals.









The following table presents the data used to create the previous charts.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Codec1\_FER1 | | | Codec1\_FER3 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| c\_1 | 70.2 | 58.8 | 64.5 | 65.2 | 52.6 | 58.9 |
| c\_2 | 85.1 | 71.9 | 78.5 | 76.2 | 57.3 | 66.8 |
| p\_1 | 66.3 | 53.8 | 60.0 | 64.2 | 50.4 | 57.3 |
| p\_2 | 72.0 | 56.7 | 64.3 | 58.9 | 44.4 | 51.7 |
| si\_1 | 69.5 | 54.1 | 61.8 | 56.3 | 41.1 | 48.7 |
| si\_2 | 78.4 | 63.2 | 70.8 | 63.3 | 48.3 | 55.8 |
| sm\_1 | 77.7 | 64.0 | 70.8 | 71.8 | 55.6 | 63.7 |
| sm\_2 | 77.4 | 65.0 | 71.2 | 68.9 | 56.1 | 62.5 |
| sp\_1 | 64.2 | 51.6 | 57.9 | 54.7 | 42.1 | 48.4 |
| sp\_2 | 85.8 | 74.0 | 79.9 | 74.4 | 60.8 | 67.6 |
| sp\_3 | 60.0 | 47.7 | 53.8 | 54.1 | 40.7 | 47.4 |
| v\_1 | 67.2 | 53.9 | 60.5 | 55.3 | 39.2 | 47.3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Codec2\_FER1 | | | Codec2\_FER3 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| c\_1 | 82.7 | 72.6 | 77.7 | 78.6 | 66.0 | 72.3 |
| c\_2 | 81.2 | 69.8 | 75.5 | 74.9 | 57.7 | 66.3 |
| p\_1 | 80.3 | 69.5 | 74.9 | 77.3 | 63.9 | 70.6 |
| p\_2 | 80.9 | 67.5 | 74.2 | 66.2 | 51.4 | 58.8 |
| si\_1 | 77.7 | 64.9 | 71.3 | 61.6 | 45.1 | 53.4 |
| si\_2 | 77.3 | 61.2 | 69.2 | 63.1 | 47.7 | 55.4 |
| sm\_1 | 84.2 | 70.1 | 77.2 | 73.5 | 58.7 | 66.1 |
| sm\_2 | 80.4 | 70.0 | 75.2 | 72.5 | 60.2 | 66.3 |
| sp\_1 | 70.0 | 56.3 | 63.2 | 62.1 | 46.8 | 54.4 |
| sp\_2 | 85.6 | 73.3 | 79.5 | 79.2 | 66.9 | 73.0 |
| sp\_3 | 81.0 | 67.6 | 74.3 | 72.8 | 56.0 | 64.4 |
| v\_1 | 70.3 | 55.5 | 62.9 | 54.0 | 38.8 | 46.4 |

### Analysis of Variance Results

The data were analyzed using Analysis of Variance techniques. The following are the overall basic results from the Analysis of Variance:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Codec | 8 | 1326938 | 165867 | 856.4 | < 2.2e-16 \*\*\* |
| SigCat | 5 | 15238 | 3048 | 15.7 | 2.50E-15 \*\*\* |
| Signal | 6 | 40742 | 6790 | 35.1 | < 2.2e-16 \*\*\* |
| Site | 1 | 109184 | 109184 | 563.7 | < 2.2e-16 \*\*\* |
| Subject | 28 | 182687 | 6525 | 33.7 | < 2.2e-16 \*\*\* |
| Codec:Signal | 40 | 36003 | 900 | 4.6 | < 2.2e-16 \*\*\* |
| Codec:Site | 8 | 20330 | 2541 | 13.1 | < 2.2e-16 \*\*\* |
| Residuals | 3125 | 605265 | 194 |  |  |

Signif. codes: 0 < \*\*\* < 0.001 < \*\* < 0.01 < \* < 0.05 < • < 0.1 < ‘ ‘ < 1

All components of the model are highly statistically significant at greater than the 99.9% level. This means that each of the aspects of the experimental design was important and rightfully included in the model, so that the effect of that component can be compensated for when analyzing the variable of interest, the difference between the codecs. However, it should be kept in mind that this experiment resulted in much data being collected, and small differences can be statistically significant, while their practical effect is minimal.

The following are the main effects (the estimated mean of each level of each variable) as determined by this analysis.

Codec main effect

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1\_FER1 | Codec1\_FER3 | Codec2\_FER1 | Codec2\_FER3 | AAC\_ FER1 | AAC\_ FER3 | hidref | lp3500 | lp7000 |
| mean | 66.2 | 56.3 | 72.9 | 62.3 | 38.7 | 33.7 | 99.8 | 31.8 | 57.2 |
| N | 358 | 358 | 358 | 358 | 358 | 358 | 358 | 358 | 358 |
| Lower Bound | 64.7 | 54.9 | 71.5 | 60.9 | 37.2 | 32.3 | 98.3 | 30.3 | 55.7 |
| Upper Bound | 67.6 | 57.8 | 74.3 | 63.7 | 40.1 | 35.2 | 101.2 | 33.2 | 58.6 |

As can be seen by comparing this table with the Pivot Table analysis means above, the two analyses give almost identical results. As mentioned, the difference between the analyses is in the width of the confidence intervals.

Signal Category main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | c | p | si | sm | sp | v |
| mean | 60.3 | 56.3 | 58.3 | 59.2 | 57.2 | 52.0 |
| N | 540 | 531 | 540 | 531 | 810 | 270 |

Although this variable is highly statistically significant, the signal categories have means that do not differ too much. The practical differences may not be too great. The statistical significance here means that the largest mean is definitely statistically significantly different from the smallest, but other differences would require a more in-depth analysis.

Codec by Signal Category (Codec:SigCat) interaction effect

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Codec | SigCat | | | | | | |
|  |  | c | p | si | sm | sp | v |
| Codec1\_FER1 | mean | 71.5 | 62.2 | 66.3 | 71.0 | 63.9 | 60.5 |
| rep | N | 60 | 59 | 60 | 59 | 90 | 30 |
| Codec1\_FER3 | mean | 62.8 | 54.5 | 52.3 | 63.1 | 54.5 | 47.3 |
| rep | N | 60 | 59 | 60 | 59 | 90 | 30 |
| Codec2\_FER1 | mean | 76.6 | 74.5 | 70.3 | 76.2 | 72.3 | 62.9 |
| rep | N | 60 | 59 | 60 | 59 | 90 | 30 |
| Codec2\_FER3 | mean | 69.3 | 64.8 | 54.4 | 66.2 | 64.0 | 46.4 |
| rep | N | 60 | 59 | 60 | 59 | 90 | 30 |
| AAC\_FER1 | mean | 39.8 | 36.4 | 44.3 | 38.4 | 36.1 | 37.9 |
| rep | N | 60 | 59 | 60 | 59 | 90 | 30 |
| AAC\_FER3 | mean | 37.6 | 31.9 | 39.3 | 32.6 | 32.3 | 24.9 |
| rep | N | 60 | 59 | 60 | 59 | 90 | 30 |
| hidref | mean | 99.8 | 99.6 | 99.8 | 100.0 | 99.7 | 100.0 |
| rep | N | 60 | 59 | 60 | 59 | 90 | 30 |
| lp3500 | mean | 32.6 | 29.7 | 33.6 | 31.1 | 32.3 | 30.3 |
| rep | N | 60 | 59 | 60 | 59 | 90 | 30 |
| lp7000 | mean | 52.9 | 52.6 | 64.7 | 54.5 | 59.5 | 57.8 |
| rep | N | 60 | 59 | 60 | 59 | 90 | 30 |

As can be seen in the above table, some codecs perform relatively better in some signal categories, while other codecs perform better in other signal categories. This is the meaning of “interaction.” The set of codec by signal category interactions above are statistically significant. Without presenting all the confidence intervals, the width of the 95% confidence intervals for the sp category is ±2.9, while the width of the 95% confidence intervals for the v category is ±5.0, and the width of the 95% confidence intervals for the other categories is ±3.6.

Signal main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | c\_1 | c\_2 | p\_1 | p\_2 | si\_1 | si\_2 |
| mean | 54.3 | 61.0 | 58.0 | 57.3 | 53.6 | 61.7 |
| N | 270 | 270 | 270 | 261 | 270 | 270 |
|  | sm\_1 | sm\_2 | sp\_1 | sp\_2 | sp\_3 | v\_1 |
| mean | 58.5 | 56.8 | 53.6 | 65.6 | 53.8 | 57.6 |
| N | 261 | 270 | 270 | 270 | 270 | 270 |

The signal main effects are shown here for completeness. The differences are statistically significant, but since the each signal is a unique item, it is not clear what use can be made of these individual means.

Site main effect

|  |  |  |
| --- | --- | --- |
|  | Ericsson | FT |
| mean | 63.4 | 51.8 |
| N | 1620 | 1602 |

The sites are statistically significantly different. Again, it is not clear what use can be made of these individual means.

Subject main effect

The subjects are statistically significantly different. The details of subject results can be found in the accompanying spreadsheets..

### Sources of variability

There is definitely a statistically significant and practically significant interaction between codecs and signals. That is, some codecs worked better for some signals than for others. These interactions can best be reviewed by studying the three charts above where, for each codec under test, the quality ratings are shown for each signal.

There is also definitely a statistically significant codec by lab interaction. In other words, some codecs performed relatively better in some testing labs than in others.

### Post-screening of data

Of the 360 sets of 7 judgments (one for each codec, reference codec, and anchor) in this experiment, 2 were eliminated by the post-screening procedure. The results of the screening procedure are coded by the Weight variable, where passing judgments received a 1 and eliminated judgments received a 0. In the pivot table, this variable can be manipulated to show the Pivot Table results with all the data. The means do not change much in a practical sense.

# Application of Selection Rules

The Selection Rules as defined in S4-(03)0837 [7] have been applied using the data collected in the experiments being analyzed here. The following are the results.

## Selection Rule 1

These rules are design criteria, and we assume for the purposes of this document that all three candidate codecs pass these rules.

## Selection Rule 2

This rule ensures that each candidate codec outperforms the better of the reference codecs in each test case. Inspecting the 3 charts above showing “all data” with confidence intervals, it is easy to verify that both candidate codecs performed better than the reference codecs. The average results from the charts above for each test case have been assembled in the following chart for easy reference.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Operating condition | AAC+ | CT | AAC | RN |
| 32 kbit/s, stereo | 75.8 | 84.9 | 38.7 | 48.0 |
| 48 kbps, stereo | 82.0 | 81.5 | 60.5 | 64.1 |
| 32 kbps, stereo, 1% FER | 66.2 | 72.9 | 38.7 | n/a |
| 32 kbps, stereo, 3% FER | 56.3 | 62.3 | 33.7 | n/a |

## Selection Rule 3

As described in the Selection Rules document, and clarified in document [9] the Preferred and Informative Figure of Merit (FoM) calculations were performed and are presented in the table below. The AAC reference is referred to as the “preferred quality FoM” and the RN reference is referred to as the “informative quality FoM.

|  |  |  |  |
| --- | --- | --- | --- |
| **AAC+** |  |  |  |
| ***Preferred FoM*** | |  |  |
|  | Mean | min | max |
| LR-A3 | 21.02 | -14.47 | 42.50 |
| LR-A4 | 6.23 | -31.70 | 35.14 |
| HR-1 | 37.05 | 22.52 | 44.24 |
| HR-2 | 21.51 | -0.44 | 47.74 |
| HR-3-1% | 27.49 | 18.10 | 37.10 |
| HR-3-3% | 12.13 | -4.37 | 28.13 |
| average | 20.90 | 13.39 | 43.03 |
| min | -31.70 |  |  |
| max | 47.74 |  |  |
| ***FoM L1*** | 6 |  |  |
| ***FoM L2*** | 0 |  |  |
|  |  |  |  |
| **CT** |  |  |  |
| ***Preferred FoM*** | |  |  |
|  | Mean | min | max |
| LR-A3 | 21.87 | -15.40 | 42.00 |
| LR-A4 | 17.91 | -13.80 | 44.24 |
| HR-1 | 46.10 | 25.62 | 70.40 |
| HR-2 | 20.99 | -2.44 | 48.07 |
| HR-3-1% | 34.22 | 16.50 | 49.17 |
| HR-3-3% | 8.71 | -15.70 | 24.10 |
| average | 24.97 | 13.23 | 55.88 |
| min | -15.70 |  |  |
| max | 70.40 |  |  |
| ***FoM L1*** | 6 |  |  |
| ***FoM L2*** | 0 |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **AAC+** |  |  |  |
| ***Informative FoM*** | |  |  |
|  | Mean | min | max |
| HR-1 | 27.84 | 4.00 | 43.57 |
| HR-2 | 18.02 | -0.15 | 33.04 |
| average | 22.93 | 1.93 | 38.30 |
| min | -0.15 |  |  |
| max | 43.57 |  |  |
| ***FoM L1*** | 2 |  |  |
| ***FoM L2*** | 0 |  |  |
|  |  |  |  |
| **CT** |  |  |  |
| ***Informative FoM*** | |  |  |
|  | Mean | min | max |
| HR-1 | 36.90 | 20.47 | 54.23 |
| HR-2 | 17.50 | -4.70 | 37.67 |
| average | 27.20 | 7.88 | 45.95 |
| min | -4.70 |  |  |
| max | 54.23 |  |  |
| ***FoM L1*** | 2 |  |  |
| ***FoM L2*** | 0 |  |  |
|  |  |  |  |

# Reference Documents

1. Tdoc S4-(03)0821, PSS/MMS High-Rate Audio Selection Test and Processing Plan Version 2.2.0.
2. Tdoc S4-(03)0824, AMR-WB+ and PSS/MMS Low-Rate Audio Selection Test and Processing Plan Version 2.2.
3. RECOMMENDATION ITU-R BS.1534, Method for the subjective assessment of intermediate quality level of coding systems.
4. An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics, Version 1.4.1, by W.N. Venables, D.M. Smith and the R Development Core Team (2001) Network Theory Limited.
5. Modern Applied Statistics with S, by W.N. Venables and B.D. Ripley (2002) Springer. Known colloquially as MASS.
6. MASS, p 140ff describe lm(). p 165ff describe aov(), which is a “wrapper” for lm().
7. Tdoc S4-(03)0837. PSS/MMS Audio Codec and Extended AMR-WB Selection Rules, Version 2.0.
8. Tdoc S4-(04)0173, Global Analysis Laboratory Report on 3GPP Low-Rate Audio Codec Exercises
9. Tdoc S4-040117 Implementation of the preferred FOM of PSS/MMS low-rate audio codec selection rule 3.