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

**Title: G****lobal Analysis Laboratory Report on 3GPP Low-Rate Audio Codec Exercises**

**Status: Approved**

**Revision: March 5, 2004**

# Executive Summary

A series of eight experiments were conducted in the 3GPP Audio codec exercise, as specified in S4-030824, “AMR-WB+ and PSS/MMS Low-Rate Audio Selection Test and Processing Plan Version 2.2.” This documents reports the results of those tests.

The following table summarizes the performance of the candidate codecs in each of the eight tests. 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 codecs at the 95% level of significance (based on ANOVA results). In the case that two codecs are “best” (e.g. test A3) it indicates that the two codecs do not differ from each other in a statistically significant sense, but that both are better than the third codec the 95% level of significance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | Operating condition | AAC+ | AMR-WB+ | CT |
| A1 | 14 kbps, mono, use case A (PSS) | 50.8 | 62.6 | 51.5 |
| A2 | 18 kbps, stereo, use case A (PSS) | 37.5 | 55.6 | 53.3 |
| A3 | 24 kbps, mono, use case A (PSS) | 75.0 | 67.4 | 75.8 |
| A4 | 24 kbps, stereo, use case A (PSS) | 55.3 | 61.3 | 67.1 |
| B1 | 14 kbps, mono, use case B (MMS),  16 kHz inp. and outp. sampling rate | 45.5 | 50.7 | 44.4 |
| B2 | 18 kbps, stereo, use case B (MMS) | 43.3 | 50.7 | 55.7 |
| B3 | 14 kbps, mono, use case A (PSS),  3% FER | 43.1 | 52.5 | 44.3 |
| B4 | 24 kbps, stereo, use case A (PSS),  3% FER | 48.9 | 53.3 | 58.0 |

As the table shows, AMR-WB+ and CT each have operating points at which they have strong performance. It appears that bit rate (i.e. lower or higher) and number of channels (i.e. mono or stereo) are significant factors in determining the performance of these two codecs.

The data support the following statements:

* In all three tests at 14 kb/s (B1, A1, B3), candidate AMR-WB+ had a mean score that was better than candidate CT in a statistical sense at the 95% confidence level.
* In one test at 18 kb/s (A2), candidate AMR-WB+ had a mean score that was better than candidate CT, while in the other test at 18 kb/s (B2), CT had a mean score that was better than AMR-WB+, where “better” is in a statistical sense at the 95% confidence level.
* In all three tests at 24 kb/s (A3, A4, B4), candidate CT had a mean score that was better than candidate AMR-WB+ in a statistical sense at the 95% confidence level.
* In all tests (A1-B4) candicate AMR-WB+ is better than reference codecs AAC and AMR-WB in a statistical sense at the 95% confidence level.
* In both tests at 18 kb/s (A2, B3), candicate CT is better than the reference codecs AAC and AMR-WB in a statistical sense at the 95% confidence level.
* In all three tests at 24 kb/s (A3, A4, B4), all candidate codecs are better than the reference codecs (AAC and AMR-WB) in a statistical sense at the 95% confidence level.

Table of Contents

[Executive Summary 1](#__RefHeading___Toc66168091)

[1 Introduction 4](#__RefHeading___Toc66168092)

[2 Overview of experiments 4](#__RefHeading___Toc66168093)

[3 Systems under test 5](#__RefHeading___Toc66168094)

[3.1 Candidate codecs 5](#__RefHeading___Toc66168095)

[3.2 Reference codecs 5](#__RefHeading___Toc66168096)

[3.3 Anchors and references 5](#__RefHeading___Toc66168097)

[4 Experimental design 6](#__RefHeading___Toc66168098)

[4.1 Experiment block A 6](#__RefHeading___Toc66168099)

[4.2 Experiment Block B 7](#__RefHeading___Toc66168100)

[5 Test Material 9](#__RefHeading___Toc66168101)

[5.1 Signal categories 9](#__RefHeading___Toc66168102)

[5.2 Training Items 9](#__RefHeading___Toc66168103)

[5.3 Test Items 9](#__RefHeading___Toc66168104)

[6 Test sites 9](#__RefHeading___Toc66168105)

[7 Statistical analysis 10](#__RefHeading___Toc66168106)

[7.1 Overview 10](#__RefHeading___Toc66168107)

[7.2 Statistical Model Based on the Experimental Design 11](#__RefHeading___Toc66168108)

[7.3 Pivot Table and ANOVA Analysis 12](#__RefHeading___Toc66168109)

[7.4 Post-Processing of Listener Data 12](#__RefHeading___Toc66168110)

[7.5 Analysis Process 12](#__RefHeading___Toc66168111)

[8 Test Results 14](#__RefHeading___Toc66168112)

[8.1 Test A1a and A1b 14](#__RefHeading___Toc66168113)

[8.2 Test A2a and A2b 21](#__RefHeading___Toc66168114)

[8.3 Test A3a and A3b 28](#__RefHeading___Toc66168115)

[8.4 Test A4a and A4b 35](#__RefHeading___Toc66168116)

[8.5 Test B1a and B1b 42](#__RefHeading___Toc66168117)

[8.6 Test B2a and B2b 49](#__RefHeading___Toc66168118)

[8.7 Test B3a and B3b 56](#__RefHeading___Toc66168119)

[8.8 Test B4a and B4b 58](#__RefHeading___Toc66168120)

[9 Application of Selection Rules 58](#__RefHeading___Toc66168121)

[9.1 PSS/MMS LBRAC Selection Rule 1 58](#__RefHeading___Toc66168122)

[9.2 PSS/MMS LBRAC Selection Rule 2 58](#__RefHeading___Toc66168123)

[9.3 PSS/MMS LBRAC Selection Rule 3 58](#__RefHeading___Toc66168124)

[10 Reference Documents 58](#__RefHeading___Toc66168125)

[Annex I - Low-Rate Experiment Training and Test Items 58](#__RefHeading___Toc66168126)

[Training Items 58](#__RefHeading___Toc66168127)

[Test Items 58](#__RefHeading___Toc66168128)

# Introduction

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

In this report, Section 2 presents an overview of the test design, Section 3 describes the systems under test, and Section 4 describes the experimental design in greater detail. Section 5 describes the test material used. For a detailed report on processing of the test material, see the Host and Mirror Laboratory Reports. Section 6 documents the test laboratories used for each component of the test. Section 7 presents an overview of the statistical analysis used in the data reduction, and Section 8 presents the test results for each of the experiments. Section 9 presents the results of applying the Selection Rules.

# Overview of experiments

There were eight experiments conducted, which were divided into two main blocks, “A” and “B”, each of which tested different operating conditions:

* A: Intrinsic quality comparison of candidate codecs
* B: Quality comparison under stressed operating conditions

Each of experiment block A and B were further divided into four experiments that tested the candidate codecs at different bitrates and operational conditions.

Experiments in block A tested the candidate codecs at the following bitrates and operating conditions:

* A1: 14 kbps, mono, use case A (PSS)
* A2: 18 kbps, stereo, use case A (PSS)
* A3: 24 kbps, mono, use case A (PSS)
* A4: 24 kbps, stereo, use case A (PSS)

Experiments in block B tested the candidate codecs at the following bitrates and operating conditions:

* B1: 14 kbps, mono, use case B (MMS), 16 kHz input and output sampling rate.
* B2: 18 kbps, stereo, use case B (MMS),
* B3: 14 kbps, mono, use case A (PSS), 3% frame error rate (FER)
* B4: 24 kbps, stereo, use case A (PSS), 3% FER

Each of experiments 1-4 in blocks A and B was further divided into two sub-experiments, designated “a” and “b”. This division made the magnitude of the resulting listening task of reasonable size and also permitted added diversity in the test material. Two listening labs participated in each sub-experiment (for a total of four per experiment), and a different set of test material was used for each sub-experiment.

* A1a Test material set A1a, Listening Lab 1 and 5
* A1b Test material set A1b, Listening Lab 2 and 6
* A2a Test material set A2a, Listening Lab 3 and 7
* A2b Test material set A2b, Listening Lab 4 and 8
* A3a Test material set A3a, Listening Lab 5 and 1
* A3b Test material set A3b, Listening Lab 6 and 2
* A4a Test material set A4a, Listening Lab 7 and 3
* A4b Test material set A4b, Listening Lab 8 and 4
* B1a Test material set B1a, Listening Lab 1 and 5
* B1b Test material set B1b, Listening Lab 2 and 6
* B2a Test material set B2a, Listening Lab 3 and 7
* B2b Test material set B2b, Listening Lab 4 and 8
* B3a Test material set B3a, Listening Lab 5 and 1
* B3b Test material set B3b, Listening Lab 6 and 2
* B4a Test material set B4a, Listening Lab 7 and 3
* B4b Test material set B4b, Listening Lab 8 and 4

# Systems under test

## Candidate codecs

The candidate codecs participating in the AMR-WB+ and PSS/MMS low-rate audio selection tests are listed in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Codec** | **AMR-WB+ candidate** | **PSS/MMS low-rate audio candidate** | **Providing Organization(s)** |
| AAC+ | No | Yes | Coding Technologies/ NEC |
| AMR-WB+ | Yes | Yes | Ericsson/  Nokia/ VoiceAge |
| CT | No | Yes | Coding Technologies |

## Reference codecs

The reference codecs are listed in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Codec name** | **AMR-WB+ candidate** | **PSS/MMS low-rate audio candidate** | **Providing Organization(s)** |
| AAC | No | No | Fraunhofer |
| AMR-WB | No | No | 3GPP |

## 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 testing mono signals, two anchors were used, those being lowpass filtered versions of the original signal. In the experiments testing stereo signals, three anchors were used, those being lowpass filtered versions of the original signal with, additionally, a reduced stereo image. The designation “side channel attenuated by 12dB” indicates that the sum and difference signals are constructed from the stereo signal, the difference signal is attenuated by 12dB, and the stereo signal is reconstructed. A similar process is followed for 6dB attenuation. One of the references is the uncoded original signal, designated the “Hidden Reference.” The other reference signal is also uncoded original signal, but it is designated the “Open Reference.” The MUSHRA test methodology [2], requires not only 3.5 kHz and 7.0 kHz Lowpass anchors, but also both Open and Hidden references.

|  |  |  |
| --- | --- | --- |
| **Type** | **Specification** | **Channels** |
| Anchor | 3.5 kHz Lowpass | Mono |
| Anchor | 7.0 kHz Lowpass | Mono |
| Anchor | 3.5 kHz Lowpass  significantly reduced stereo image  (side channel attenuated by 12dB) | Stereo |
| Anchor | 7.0 kHz Lowpass  significantly reduced stereo image  (side channel attenuated by 12dB) | Stereo |
| Anchor | 7.0 kHz Lowpass  slightly reduced stereo image  (side channel attenuated by 6dB) | Stereo |
| Hidden Reference | Original signal | Mono and Stereo |
| 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 (composed of two sub-experiments). Each experiment is specified in a separate table.
* The row labeled “Bit Rate” indicates the bitrate for the experiment. All candidate and reference codecs run at this bitrate unless explicitly noted in the “Additional Constraints” column (e.g. as with AMR-WB in experiment A1a and A1b).
* The row labeled “Signal” indicates the number of distinct channels in the test material (i.e. mono or stereo). All signals are 48 kHz sampling rate unless explicitly noted in the “Additional Constraints” column. If noted in the “Signal” row (e.g. as in experiment B1a and B1b) this indicates that all codecs processed a sampling rate other than 48 kHz. If indicated in a “codec” row (e.g. as with AMR-WB in experiment A1a and A1b), it indicates that that codec processed a sampling rate other than 48 kHz.
* 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.
* The row labeled “Anchors and references” lists each anchor and reference condition tested in the experiment in sub-divisions of the main row.

## Experiment block A

All experiments in block A are use case A (PSS) and the test material used in each experiment is described in Section 5.

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Additional Constraints** |
| Experiment | A1a and A1b |  |
| Bit Rate | 14 kbps |  |
| Signal | Mono |  |
| Candidate codecs | AAC+ |  |
| AMR-WB+ |  |
| CT |  |
| Reference codecs | AAC |  |
| AMR-WB | 14.25 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 | A2a and A2b |  |
| Bit Rate | 18 kbps |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+ |  |
| AMR-WB+ |  |
| CT |  |
| Reference codecs | AAC |  |
| AMR-WB | 18.25 kbps, 16 kHz sampling rate, mono |
| Anchors and references | Open Reference |  |
| Hidden Reference |  |
| 7.0 kHz Lowpass | 6 dB attenuated side channel |
| 7.0 kHz Lowpass | 12 dB attenuated side channel |
| 3.5 kHz Lowpass | 12 dB attenuated side channel |

|  |  |  |
| --- | --- | --- |
| **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, mono |
| 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 |

## Experiment Block B

In block B, experiments B1a, B1b, B2a and B2b are use case B (MMS) while experiments B3a, B3b, B4a and B4b are use case A (PSS). Test Material used in each experiment is described in Section 5. This table for Experiments B3a, B3b, B4a and B4b have a new row that indicates “Channel Error Condition.” These experiments are tested under simulated errored channel conditions, such that on average three percent of the codec frames are errored.

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Additional Constraints** |
| Experiment | B1a and B1b |  |
| Bit Rate | 14 kbps |  |
| Signal | Mono | 16 kHz input and output sampling rate |
| Candidate codecs | AAC+ |  |
| AMR-WB+ |  |
| CT |  |
| Reference codecs | AAC |  |
| AMR-WB | 14.25 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 | B2a and B2b |  |
| Bit Rate | 18 kbps |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+ |  |
| AMR-WB+ |  |
| CT |  |
| Reference codecs | AAC |  |
| AMR-WB | 18.25 kbps, 16 kHz sampling rate, mono |
| Anchors and references | Open Reference |  |
| Hidden Reference |  |
| 7.0 kHz Lowpass | 6 dB attenuated side channel |
| 7.0 kHz Lowpass | 12 dB attenuated side channel |
| 3.5 kHz Lowpass | 12 dB attenuated side channel |

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Additional Constraints** |
| Experiment | B3a and B3b |  |
| Bit Rate | 14 kbps |  |
| Signal | Mono |  |
| Channel Error Condition | 3% FER |  |
| Candidate codecs | AAC+ |  |
| AMR-WB+ |  |
| CT |  |
| Reference codecs | AAC |  |
| AMR-WB | 14.25 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 | B4a and B4b |  |
| Bit Rate | 24 kbps |  |
| Signal | Stereo |  |
| Channel Error Condition | 3% FER |  |
| Candidate codecs | AAC+ |  |
| AMR-WB+ |  |
| CT |  |
| Reference codecs | AAC |  |
| AMR-WB | 23.85 kbps, 16 kHz sampling rate, mono |
| Anchors and references | Open Reference |  |
| Hidden Reference |  |
| 7.0 kHz Lowpass | 6 dB attenuated side channel |
| 7.0 kHz Lowpass | 12 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 four signal categories:

* Music
* Speech
* Speech over music (i.e. speech with background music)
* Speech between music (i.e. alternating speech and music segments)

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

## Training Items

A single set of four training items were used for the eight tests, one item selected from each of the four stimulus categories. The four training items are shown in Annex I.

## Test Items

Eight sets of test items were used, one for each experiment. The four signal categories were represented within each set, specifically with four Music items, four Speech items, two Speech between Music items and two Speech over Music items. Due to limitations in the availability of test material, some individual items appeared in more than one set. The eight sets are shown in Annex I.

# Test sites

Individual experiments use two listening laboratories, as shown in Table 7-1. The abbreviation for the listening labs are as follows: Fraunhofer Geselschaft (FhG), France Telecom (FT) , T-Systems (TS), NTT-AT, Dynastat (D), Nokia (N), Ericsson (E), Coding Technologies (CT).

Table 7‑1: Allocation of sub-experiments to the Listening Laboratories

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Exp. | Lab1 | Lab2 | Lab3 | Lab4 | Lab5 | Lab6 | Lab7 | Lab8 |
| LL ID | FhG | CT | E | N | D | FT | TS | NTT\_AT |
| A1a | x |  |  |  | X |  |  |  |
| A1b |  | x |  |  |  | X |  |  |
| A2a |  |  | x |  |  |  | x |  |
| A2b |  |  |  | x |  |  |  | x |
| A3a | x |  |  |  | X |  |  |  |
| A3b |  | x |  |  |  | X |  |  |
| A4a |  |  | x |  |  |  | x |  |
| A4b |  |  |  | x |  |  |  | x |
| B1a | x |  |  |  | X |  |  |  |
| B1b |  | x |  |  |  | X |  |  |
| B2a |  |  | x |  |  |  | x |  |
| B2b |  |  |  | x |  |  |  | x |
| B3a | x |  |  |  | X |  |  |  |
| B3b |  | x |  |  |  | X |  |  |
| B4a |  |  | x |  |  |  | x |  |
| B4b |  |  |  | x |  |  |  | x |
| Totals: | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |

# Statistical analysis

## Overview

### Standard Pivot Table Analysis

The Pivot Table statistical analysis followed the standard MUSHRA procedure [2].

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

The mean score , is calculated as:

where:

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

is the weight for test sequence *k*

Note that in this test, signal categories Speech over Music and Speech between Music had a weight of 2, with all other categories having a weight of 1.

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:

where *N* is the number of independent observations (typically number of observers times number of sequences) 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.

## 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 sub-experiment 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. Similarly, laboratories are nested within sub-experiments for the low-rate experiments. And, for the low-rate experiments, audio selections are also nested within sub-experiments.

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

Score = Codec (c = 1, …, 8 or 9)

+ Sub-experiment (Sub = a or b)

+ Signal Category (SigCat = 1, … 4)

+ Signal (Signal = 1, …, 24)

+ Codec by Signal Category interaction

(Codec:SigCat, Codec = 1, …, 8 or 9, SigCat = 1, …, 4)

+ Laboratory (Site = 1, …, 4)

+ Codec by Laboratory interaction (Codec:Site, Codec = 1, …, 8 or 9, Site = 1, …, 4)

+ 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 8 low-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 9.

## 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 95 rather than exactly 100. Similarly, scores may depart from strict monotonicity by 5 points and still be allowed.

## 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 multiple 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 95 or
   * the lp3500 anchor rating is not more than 5 units greater than the lp7000 anchor rating. In some experiments there are two lp3500 anchors or two lp700 anchors. In those cases, the two ratings are averaged before the comparison.
7. Data is weighted according to the “relative values” given in Table 3.2 of [7].
8. 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.
9. The data is exported to a text file and entered into “R” [3], a gnu version of the statistical analysis system called “S” [4]. A script is used to fit the model. In particular, the function aov() [5]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.
10. 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, Codec2 and Codec3 such that their identity is concealed.

## Test A1a and A1b

### Test parameters and systems under test

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Symbol** |
| Experiment | A1a and A1b |  |
| Bit Rate | 14 kbps |  |
| Signal | Mono |  |
| Candidate codecs | AAC+ | Codec 1 |
| AMR-WB+ | Codec 2 |
| CT | Codec 3 |
| Reference codecs | AAC | AAC |
| AMR-WB, 14.25 kbps, 16 kHz sampling rate | AMR-WB |
| 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 candidates codecs out-performs both of the reference codecs. The following table shows the numerical values plotted in the chart above.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | Codec3 | AAC | AMR-WB | hidref | lp3500 | lp7000 |
| Average | 50.8 | 62.6 | 51.5 | 32.9 | 44.9 | 99.9 | 29.5 | 62.5 |
| Lower Bound | 48.9 | 60.9 | 49.7 | 31.3 | 43.1 | 99.9 | 28.1 | 61.0 |
| Upper Bound | 52.8 | 64.4 | 53.4 | 34.5 | 46.8 | 100.0 | 30.9 | 64.1 |

The following 3 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 | | | Codec 3 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| m\_ot\_x\_8 | 72.4 | 57.4 | 64.9 | 71.3 | 59.4 | 65.3 | 68.1 | 54.6 | 61.4 |
| m\_ot\_x\_9 | 58.7 | 45.2 | 51.9 | 68.0 | 55.6 | 61.8 | 60.8 | 46.9 | 53.8 |
| m\_ot\_x\_a | 80.7 | 67.5 | 74.1 | 73.6 | 60.0 | 66.8 | 76.1 | 64.4 | 70.2 |
| m\_ot\_x\_b | 61.7 | 49.1 | 55.4 | 45.7 | 33.8 | 39.8 | 58.8 | 45.9 | 52.4 |
| m\_po\_x\_5 | 75.3 | 61.1 | 68.2 | 73.2 | 62.5 | 67.9 | 71.7 | 57.8 | 64.8 |
| m\_po\_x\_6 | 58.8 | 47.1 | 52.9 | 51.4 | 38.6 | 45.0 | 58.9 | 48.3 | 53.6 |
| m\_po\_x\_7 | 81.3 | 67.7 | 74.5 | 75.1 | 61.5 | 68.3 | 79.1 | 68.0 | 73.6 |
| m\_si\_x\_3 | 53.7 | 38.2 | 45.9 | 77.0 | 63.9 | 70.5 | 57.4 | 40.4 | 48.9 |
| s\_cl\_2t\_3 | 47.2 | 32.6 | 39.9 | 88.5 | 80.3 | 84.4 | 52.5 | 37.1 | 44.8 |
| s\_cl\_2t\_4 | 47.8 | 34.9 | 41.3 | 83.8 | 72.7 | 78.3 | 52.1 | 37.9 | 45.0 |
| s\_cl\_2t\_5 | 53.8 | 40.0 | 46.9 | 66.4 | 55.2 | 60.8 | 54.3 | 41.8 | 48.1 |
| s\_cl\_mt\_2 | 29.8 | 17.3 | 23.6 | 79.3 | 69.5 | 74.4 | 30.3 | 18.5 | 24.4 |
| s\_no\_2t\_1 | 77.4 | 62.4 | 69.9 | 80.9 | 70.9 | 75.9 | 73.2 | 59.3 | 66.2 |
| s\_no\_2t\_2 | 57.0 | 42.7 | 49.8 | 62.6 | 49.8 | 56.2 | 56.5 | 42.8 | 49.6 |
| s\_no\_ft\_1 | 64.4 | 51.9 | 58.1 | 72.5 | 59.3 | 65.9 | 66.0 | 53.8 | 59.9 |
| s\_no\_ft\_2 | 44.3 | 32.0 | 38.2 | 75.0 | 65.7 | 70.3 | 50.1 | 38.8 | 44.5 |
| sbm\_js\_x\_1 | 41.9 | 27.7 | 34.8 | 77.5 | 63.2 | 70.3 | 42.3 | 29.4 | 35.9 |
| sbm\_ms\_x\_1 | 49.0 | 35.0 | 42.0 | 81.4 | 67.4 | 74.4 | 51.1 | 36.7 | 43.9 |
| sbm\_sj\_x\_1 | 47.6 | 33.8 | 40.7 | 60.6 | 47.6 | 54.1 | 52.7 | 39.5 | 46.1 |
| sbm\_sm\_x\_6 | 59.2 | 46.2 | 52.7 | 74.3 | 65.7 | 70.0 | 58.1 | 44.8 | 51.5 |
| som\_fi\_x\_4 | 56.4 | 42.1 | 49.3 | 70.7 | 58.4 | 64.5 | 54.6 | 41.0 | 47.8 |
| som\_ot\_x\_4 | 74.0 | 59.1 | 66.6 | 68.5 | 52.6 | 60.6 | 72.5 | 56.9 | 64.7 |
| som\_ot\_x\_5 | 66.4 | 51.0 | 58.7 | 56.9 | 43.3 | 50.1 | 67.4 | 52.9 | 60.2 |
| som\_ot\_x\_6 | 50.2 | 33.7 | 42.0 | 43.4 | 28.4 | 35.9 | 52.9 | 37.2 | 45.1 |

### 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 | 7 | 3176083 | 453726 | 2213.8 | < 2.2e-16 \*\*\* |
| Sub | 1 | 174779 | 174779 | 852.8 | < 2.2e-16 \*\*\* |
| SigCat | 3 | 8119 | 2706 | 13.2 | 1.36e-08 \*\*\* |
| Signal | 19 | 49926 | 2628 | 12.8 | < 2.2e-16 \*\*\* |
| Site | 2 | 140772 | 70386 | 343.4 | < 2.2e-16 \*\*\* |
| Subject | 57 | 435998 | 7649 | 37.3 | < 2.2e-16 \*\*\* |
| Codec:Signal | 21 | 227576 | 10837 | 52.9 | < 2.2e-16 \*\*\* |
| Codec:Site | 21 | 95383 | 4542 | 22.2 | < 2.2e-16 \*\*\* |
| Residuals | 7340 | 1504370 | 205 |  |  |

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 | Codec3 | AAC | AMR-WB | hidref | lp3500 | lp7000 |
| mean | 50.8 | 62.7 | 51.6 | 32.9 | 45.0 | 99.9 | 29.5 | 62.6 |
| N | 701 | 701 | 701 | 701 | 701 | 701 | 701 | 701 |
| Lower Bound | 49.8 | 61.6 | 50.5 | 31.9 | 43.9 | 98.9 | 28.4 | 61.5 |
| Upper Bound | 51.9 | 63.7 | 52.6 | 34.0 | 46.0 | 101.0 | 30.5 | 63.6 |

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.

Sub Experiment main effect

|  |  |  |
| --- | --- | --- |
|  | a | b |
| mean | 59.4 | 49.7 |
| N | 2706 | 2898 |

The two sub experiments have surprisingly different means. This difference is not only statistically significant, it may lead to insight about the differences between the signal sets or the laboratories employed in the two sub-experiments.

Signal Category main effect

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m | s | sbm | som |
| mean | 54.8 | 55.8 | 53.1 | 53.7 |
| N | 1896 | 1896 | 920 | 920 |

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 | | | | |
|  |  | m | s | sbm | som |
| Codec1 | mean | 60.8 | 45.6 | 42.8 | 53.9 |
|  | N | 237 | 237 | 115 | 115 |
| Codec2 | mean | 60.3 | 70.7 | 67.1 | 52.4 |
|  | N | 237 | 237 | 115 | 115 |
| Codec3 | mean | 59.7 | 47.5 | 44.6 | 54.3 |
|  | N | 237 | 237 | 115 | 115 |
| AAC | mean | 40.8 | 30.1 | 28.0 | 32.7 |
|  | N | 237 | 237 | 115 | 115 |
| AMR-WB | mean | 31.0 | 55.6 | 50.3 | 43.0 |
|  | N | 237 | 237 | 115 | 115 |
| hidref | mean | 99.8 | 100.0 | 100.0 | 100.0 |
|  | N | 237 | 237 | 115 | 115 |
| lp3500 | mean | 27.2 | 32.2 | 29.3 | 29.2 |
|  | N | 237 | 237 | 115 | 115 |
| lp7000 | mean | 58.7 | 65.0 | 62.7 | 64.0 |
|  | N | 237 | 237 | 115 | 115 |

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 m and s categories is ±1.8, while the width of the 95% confidence intervals for the som and sbm categories is ±2.6.

Signal main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | m\_ot\_x\_8 | m\_ot\_x\_9 | m\_ot\_x\_a | m\_ot\_x\_b | m\_po\_x\_5 | m\_po\_x\_6 |
| mean | 55.1 | 55.8 | 56.8 | 49.4 | 53.6 | 51.2 |
| N | 224 | 232 | 240 | 248 | 240 | 248 |
|  | m\_po\_x\_7 | m\_si\_x\_3 | s\_cl\_2t\_3 | s\_cl\_2t\_4 | s\_cl\_2t\_5 | s\_cl\_mt\_2 |
| mean | 56.5 | 57.1 | 52.5 | 54.3 | 55.4 | 53.1 |
| N | 232 | 232 | 232 | 232 | 240 | 248 |
|  | s\_no\_2t\_1 | s\_no\_2t\_2 | s\_no\_ft\_1 | s\_no\_ft\_2 | sbm\_js\_x\_1 | sbm\_ms\_x\_1 |
| mean | 59.1 | 53.4 | 52.5 | 54.8 | 51.4 | 52.3 |
| N | 224 | 248 | 224 | 248 | 216 | 224 |
|  | sbm\_sj\_x\_1 | sbm\_sm\_x\_6 | som\_fi\_x\_4 | som\_ot\_x\_4 | som\_ot\_x\_5 | som\_ot\_x\_6 |
| mean | 54.5 | 58.8 | 52.6 | 58.4 | 55.5 | 51.2 |
| N | 240 | 240 | 224 | 216 | 240 | 240 |

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CT | DY | FhG | FT |
| mean | 56.3 | 60.6 | 48.9 | 52.1 |
| N | 1536 | 1266 | 1440 | 1362 |

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. However, the effect of this interaction compared to, say, the listener differences, the signal differences or the codec-signal interaction is relatively small. If this interaction had not been included in the statistical model, the residual standard error would have been about 4% larger.

### Post-screening of data

Of the 732 sets of 8 judgments (one for each codec, reference codec, and anchor) in this experiment, 28 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. However, in the analysis of variance, the standard error of the residuals, and thus all confidence interval widths, increases by about 5%.

## Test A2a and A2b

### Test parameters and systems under test

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Symbol** |
| Experiment | A2a and A2b |  |
| Bit Rate | 18 kbps |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+ | Codec 1 |
| AMR-WB+ | Codec 2 |
| CT | Codec 3 |
| Reference codecs | AAC | AAC |
| AMR-WB, 18.25 kbps, 16 kHz sampling rate | AMR-WB |
| Anchors and references | Open Reference |  |
| Hidden Reference | HR |
| 7.0 kHz Lowpass, 6 dB attenuated side channel | LP7.0-S6 |
| 7.0 kHz Lowpass,12 dB attenuated side channel | LP7.0-S12 |
| 3.5 kHz Lowpass, 12 dB attenuated side channel | LP3.5-S12 |

### 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.



Codec1 did not perform better than AMR-WB. The other two candidate codecs out-perform both of the reference codecs. The following table shows the numerical values plotted in the chart above.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | Codec3 | AAC | AMR-WB | hidref | lp3500\_s12 | lp7000\_s12 | lp7000\_s6 |
| Average | 37.5 | 55.6 | 53.3 | 20.9 | 48.2 | 100.0 | 31.2 | 60.6 | 62.3 |
| Lower Bound | 35.5 | 53.4 | 50.8 | 19.4 | 46.4 | 99.9 | 29.6 | 58.9 | 60.6 |
| Upper Bound | 39.5 | 57.8 | 55.7 | 22.4 | 50.1 | 100.0 | 32.8 | 62.3 | 64.0 |

The following 3 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 | | | Codec 3 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| m\_ot\_x\_4 | 44.1 | 30.1 | 37.1 | 59.9 | 42.9 | 51.4 | 78.7 | 65.9 | 72.3 |
| m\_ot\_x\_5 | 54.7 | 37.5 | 46.1 | 57.8 | 40.9 | 49.3 | 81.7 | 69.2 | 75.4 |
| m\_ot\_x\_6 | 71.9 | 59.4 | 65.7 | 65.1 | 47.2 | 56.1 | 80.7 | 62.6 | 71.6 |
| m\_ot\_x\_7 | 63.3 | 44.4 | 53.8 | 80.2 | 65.4 | 72.8 | 84.0 | 67.6 | 75.8 |
| m\_po\_x\_2 | 42.2 | 28.7 | 35.5 | 50.5 | 35.5 | 43.0 | 69.0 | 53.7 | 61.4 |
| m\_po\_x\_3 | 53.2 | 36.6 | 44.9 | 41.9 | 27.1 | 34.5 | 73.3 | 57.7 | 65.5 |
| m\_po\_x\_4 | 53.0 | 38.7 | 45.8 | 59.0 | 42.8 | 50.9 | 59.8 | 44.0 | 51.9 |
| m\_si\_x\_2 | 59.3 | 43.1 | 51.2 | 67.7 | 49.8 | 58.7 | 78.5 | 63.3 | 70.9 |
| s\_cl\_2t\_4 | 28.4 | 16.8 | 22.6 | 76.7 | 60.3 | 68.5 | 36.1 | 23.2 | 29.6 |
| s\_cl\_2t\_5 | 56.9 | 40.8 | 48.8 | 74.1 | 54.6 | 64.4 | 74.0 | 55.1 | 64.5 |
| s\_cl\_ft\_3 | 28.1 | 16.0 | 22.0 | 76.1 | 58.8 | 67.4 | 50.5 | 32.2 | 41.4 |
| s\_cl\_mt\_2 | 30.1 | 19.3 | 24.7 | 83.4 | 64.2 | 73.8 | 39.5 | 25.0 | 32.3 |
| s\_no\_2t\_3 | 27.2 | 16.6 | 21.9 | 61.7 | 44.3 | 53.0 | 36.9 | 23.3 | 30.1 |
| s\_no\_ft\_3 | 60.3 | 42.8 | 51.6 | 55.6 | 35.6 | 45.6 | 53.5 | 35.4 | 44.5 |
| s\_no\_ft\_4 | 40.4 | 27.8 | 34.1 | 66.4 | 47.3 | 56.8 | 43.3 | 28.4 | 35.9 |
| s\_no\_mt\_1 | 33.2 | 17.8 | 25.5 | 71.2 | 50.4 | 60.8 | 41.4 | 24.6 | 33.0 |
| sbm\_js\_x\_1 | 25.3 | 14.6 | 19.9 | 53.0 | 34.7 | 43.9 | 30.4 | 18.5 | 24.4 |
| sbm\_js\_x\_2 | 43.6 | 30.1 | 36.9 | 77.3 | 61.3 | 69.3 | 59.4 | 39.4 | 49.4 |
| sbm\_sm\_x\_4 | 24.1 | 15.2 | 19.7 | 67.1 | 51.2 | 59.1 | 51.3 | 34.2 | 42.8 |
| sbm\_sm\_x\_5 | 51.4 | 34.6 | 43.0 | 75.5 | 62.2 | 68.9 | 88.8 | 77.6 | 83.2 |
| som\_ad\_x\_1 | 49.7 | 32.7 | 41.2 | 66.7 | 49.3 | 58.0 | 64.2 | 46.0 | 55.1 |
| som\_fi\_x\_3 | 49.1 | 33.1 | 41.1 | 56.8 | 39.8 | 48.3 | 74.4 | 60.3 | 67.4 |
| som\_ot\_x\_2 | 32.5 | 18.8 | 25.6 | 43.1 | 27.2 | 35.2 | 36.1 | 21.1 | 28.6 |
| som\_ot\_x\_3 | 70.5 | 55.2 | 62.8 | 66.0 | 46.8 | 56.4 | 85.6 | 72.1 | 78.8 |

### 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 | 3800935 | 475117 | 2281.9 | < 2.2e-16 \*\*\* |
| Sub | 1 | 197314 | 197314 | 947.7 | < 2.2e-16 \*\*\* |
| SigCat | 3 | 1212 | 404 | 1.9 | 0.1207 |
| Signal | 19 | 35863 | 1888 | 9.1 | < 2.2e-16 \*\*\* |
| Site | 2 | 653224 | 326612 | 1568.6 | < 2.2e-16 \*\*\* |
| Subject | 56 | 541981 | 9678 | 46.5 | < 2.2e-16 \*\*\* |
| Codec:Signal | 24 | 295401 | 12308 | 59.1 | < 2.2e-16 \*\*\* |
| Codec:Site | 24 | 151884 | 6329 | 30.4 | < 2.2e-16 \*\*\* |
| Residuals | 8169 | 1700890 | 208 |  |  |

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 signal category (SigCat). This means that each of the aspects of the experimental design, except SigCat 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 | Codec3 | AAC | AMR-WB | hidref | lp3500\_s12 | lp7000\_s12 | lp7000\_s6 |
| mean | 37.5 | 55.6 | 53.3 | 20.9 | 48.2 | 100.0 | 31.2 | 60.6 | 62.3 |
| N | 695 | 695 | 695 | 695 | 695 | 695 | 695 | 695 | 695 |
| Lower Bound | 36.4 | 54.6 | 52.2 | 19.8 | 47.1 | 98.9 | 30.1 | 59.5 | 61.2 |
| Upper Bound | 38.6 | 56.7 | 54.4 | 22.0 | 49.3 | 101.0 | 32.3 | 61.7 | 63.3 |

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.

Sub Experiment main effect

|  |  |  |
| --- | --- | --- |
|  | a | b |
| mean | 47.5 | 57.2 |
| N | 3231 | 3024 |

The two sub experiments have surprisingly different means. This difference is not only statistically significant, it may lead to insight about the differences between the signal sets or the laboratories employed in the two sub-experiments.

Signal Category main effect

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m | s | sbm | som |
| mean | 52.2 | 52.5 | 51.6 | 52.5 |
| N | 2124 | 2079 | 1044 | 1008 |

This variable is not statistically significant. The signal categories have means that do not differ statistically.

Codec by Signal Category (Codec:SigCat) interaction effect

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Codec | SigCat | | | | |
|  |  | m | s | sbm | som |
| Codec1 | mean | 47.1 | 31.2 | 29.5 | 42.3 |
|  | N | 236 | 231 | 116 | 112 |
| Codec2 | mean | 51.9 | 61.3 | 60.0 | 49.2 |
|  | N | 236 | 231 | 116 | 112 |
| Codec3 | mean | 68.0 | 38.8 | 49.1 | 57.1 |
|  | N | 236 | 231 | 116 | 112 |
| AAC | mean | 27.6 | 15.1 | 19.1 | 21.7 |
|  | N | 236 | 231 | 116 | 112 |
| AMR-WB | mean | 36.6 | 58.4 | 50.4 | 47.7 |
|  | N | 236 | 231 | 116 | 112 |
| hidref | mean | 99.9 | 100.0 | 100.0 | 100.0 |
|  | N | 236 | 231 | 116 | 112 |
| lp3500\_s12 | mean | 28.5 | 34.4 | 32.2 | 29.7 |
|  | N | 236 | 231 | 116 | 112 |
| lp7000\_s12 | mean | 54.0 | 65.7 | 61.1 | 61.7 |
|  | N | 236 | 231 | 116 | 112 |
| lp7000\_s6 | mean | 55.7 | 67.5 | 62.8 | 63.3 |
|  | N | 236 | 231 | 116 | 112 |

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 m and s categories is ±1.9, while the width of the 95% confidence intervals for the som and sbm categories is ±2.6.

Signal main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | m\_ot\_x\_4 | m\_ot\_x\_5 | m\_ot\_x\_6 | m\_ot\_x\_7 | m\_po\_x\_2 | m\_po\_x\_3 |
| mean | 52.8 | 54.4 | 54.7 | 56.8 | 48.0 | 51.6 |
| N | 270 | 270 | 234 | 270 | 270 | 270 |
|  | m\_po\_x\_4 | m\_si\_x\_2 | s\_cl\_2t\_4 | s\_cl\_2t\_5 | s\_cl\_ft\_3 | s\_cl\_mt\_2 |
| mean | 49.5 | 50.0 | 52.2 | 54.2 | 55.0 | 51.1 |
| N | 270 | 270 | 270 | 252 | 270 | 243 |
|  | s\_no\_2t\_3 | s\_no\_ft\_3 | s\_no\_ft\_4 | s\_no\_mt\_1 | sbm\_js\_x\_1 | sbm\_js\_x\_2 |
| mean | 51.9 | 49.0 | 51.7 | 52.1 | 50.9 | 50.6 |
| N | 270 | 252 | 252 | 270 | 270 | 261 |
|  | sbm\_sm\_x\_4 | sbm\_sm\_x\_5 | som\_ad\_x\_1 | som\_fi\_x\_3 | som\_ot\_x\_2 | som\_ot\_x\_3 |
| mean | 52.1 | 55.3 | 49.9 | 54.1 | 51.2 | 53.4 |
| N | 270 | 243 | 243 | 270 | 261 | 234 |

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 | Nokia | NTT-AT | T-Sys |
| mean | 63.9 | 56.0 | 48.0 | 40.5 |
| N | 1611 | 1584 | 1440 | 1620 |

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. However, the effect of this interaction compared to, say, the listener differences, the signal differences or the codec-signal interaction is relatively small. If this interaction had not been included in the statistical model, the residual standard error would have been about 9% larger.

### Post-screening of data

Of the 720 sets of 9 judgments (one for each codec, reference codec, and anchor) in this experiment, 25 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. However, in the analysis of variance, the standard error of the residuals, and thus all confidence interval widths, increases by about 2%.

## Test A3a and A3b

### Test parameters and systems under test

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Symbol** |
| Experiment | A3a and A3b |  |
| Bit Rate | 24 kbps |  |
| Signal | Mono |  |
| Candidate codecs | AAC+ | Codec 1 |
| AMR-WB+ | Codec 2 |
| CT | Codec 3 |
| Reference codecs | AAC | AAC |
| AMR-WB, 23.85 kbps, 16 kHz sampling rate | AMR-WB |
| Anchors and references | Open Reference |  |
| Hidden Reference | HR |
| 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 candidates codecs out-performs both of the reference codecs. The following table shows the numerical values plotted in the chart above.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | Codec3 | AAC | AMR-WB | hidref | lp3500 | lp7000 |
| Average | 74.9 | 67.4 | 75.8 | 50.9 | 47.4 | 99.9 | 28.6 | 56.2 |
| Lower Bound | 73.2 | 65.6 | 74.1 | 49.0 | 45.7 | 99.9 | 27.3 | 54.7 |
| Upper Bound | 76.7 | 69.1 | 77.5 | 52.7 | 49.0 | 100.0 | 29.9 | 57.8 |

The following 3 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 | | | Codec 3 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| m\_ot\_x\_2 | 96.8 | 91.6 | 94.2 | 80.4 | 68.9 | 74.7 | 94.6 | 87.1 | 90.9 |
| m\_ot\_x\_3 | 84.8 | 72.8 | 78.8 | 52.9 | 41.8 | 47.3 | 84.8 | 74.3 | 79.5 |
| m\_po\_x\_1 | 87.7 | 78.7 | 83.2 | 72.3 | 60.0 | 66.2 | 89.4 | 78.1 | 83.7 |
| m\_po\_x\_2 | 96.7 | 88.6 | 92.7 | 68.0 | 57.9 | 63.0 | 93.2 | 86.8 | 90.0 |
| m\_po\_x\_3 | 83.5 | 72.5 | 78.0 | 51.2 | 39.1 | 45.2 | 87.6 | 77.9 | 82.7 |
| m\_po\_x\_4 | 65.9 | 49.9 | 57.9 | 59.8 | 50.1 | 55.0 | 70.2 | 56.2 | 63.2 |
| m\_si\_x\_1 | 62.6 | 48.2 | 55.4 | 76.1 | 59.9 | 68.0 | 62.3 | 47.7 | 55.0 |
| m\_si\_x\_2 | 81.6 | 70.4 | 76.0 | 55.5 | 43.9 | 49.7 | 79.6 | 66.6 | 73.1 |
| s\_cl\_2t\_1 | 73.6 | 61.3 | 67.5 | 89.7 | 80.1 | 84.9 | 75.3 | 62.7 | 69.0 |
| s\_cl\_2t\_2 | 68.5 | 55.5 | 62.0 | 72.5 | 59.8 | 66.1 | 78.2 | 66.6 | 72.4 |
| s\_cl\_ft\_3 | 78.2 | 62.7 | 70.4 | 93.2 | 83.8 | 88.5 | 79.6 | 66.5 | 73.0 |
| s\_cl\_mt\_2 | 50.0 | 34.2 | 42.1 | 85.2 | 76.3 | 80.8 | 48.9 | 33.5 | 41.2 |
| s\_no\_2t\_1 | 94.7 | 87.3 | 91.0 | 85.9 | 75.9 | 80.9 | 94.8 | 87.8 | 91.3 |
| s\_no\_2t\_2 | 83.6 | 67.4 | 75.5 | 68.5 | 56.5 | 62.5 | 84.7 | 72.7 | 78.7 |
| s\_no\_ft\_4 | 72.2 | 56.5 | 64.3 | 85.7 | 74.0 | 79.8 | 73.4 | 58.3 | 65.9 |
| s\_no\_mt\_1 | 86.3 | 72.9 | 79.6 | 94.2 | 85.6 | 89.9 | 88.0 | 75.7 | 81.8 |
| sbm\_js\_x\_2 | 73.6 | 61.8 | 67.7 | 64.4 | 51.6 | 58.0 | 76.0 | 62.8 | 69.4 |
| sbm\_ms\_x\_1 | 69.0 | 55.6 | 62.3 | 83.7 | 71.3 | 77.5 | 71.9 | 57.3 | 64.6 |
| sbm\_sm\_x\_2 | 93.9 | 77.7 | 85.8 | 82.1 | 67.1 | 74.6 | 93.5 | 77.4 | 85.4 |
| sbm\_sm\_x\_5 | 90.5 | 81.2 | 85.9 | 59.2 | 46.5 | 52.9 | 88.9 | 80.0 | 84.4 |
| som\_nt\_x\_1 | 95.4 | 88.5 | 92.0 | 78.5 | 69.8 | 74.1 | 95.5 | 88.6 | 92.0 |
| som\_ot\_x\_1 | 77.5 | 63.6 | 70.5 | 60.2 | 48.6 | 54.4 | 77.3 | 61.2 | 69.3 |
| som\_ot\_x\_2 | 78.3 | 65.1 | 71.7 | 82.2 | 70.8 | 76.5 | 79.2 | 65.9 | 72.6 |
| som\_ot\_x\_3 | 86.9 | 77.1 | 82.0 | 65.7 | 52.2 | 59.0 | 86.3 | 77.7 | 82.0 |

### 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 | 7 | 3099551 | 442793 | 2411.5 | < 2.2e-16 \*\*\* |
| Sub | 1 | 243230 | 243230 | 1324.6 | < 2.2e-16 \*\*\* |
| SigCat | 3 | 40927 | 13642 | 74.3 | < 2.2e-16 \*\*\* |
| Signal | 19 | 87161 | 4587 | 25.0 | < 2.2e-16 \*\*\* |
| Site | 2 | 194847 | 97423 | 530.6 | < 2.2e-16 \*\*\* |
| Subject | 56 | 312996 | 5589 | 30.4 | < 2.2e-16 \*\*\* |
| Codec:Signal | 21 | 106306 | 5062 | 27.6 | < 2.2e-16 \*\*\* |
| Codec:Site | 21 | 145639 | 6935 | 37.8 | < 2.2e-16 \*\*\* |
| Residuals | 7357 | 1350893 | 184 |  |  |

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 | Codec3 | AAC | AMR-WB | hidref | lp3500 | lp7000 |
| mean | 75.0 | 67.4 | 75.8 | 50.9 | 47.4 | 99.9 | 28.6 | 56.2 |
| N | 703 | 703 | 703 | 703 | 703 | 703 | 703 | 703 |
| Lower Bound | 73.9 | 66.4 | 74.8 | 49.9 | 46.4 | 98.9 | 27.6 | 55.2 |
| Upper Bound | 76.0 | 68.4 | 76.8 | 51.9 | 48.4 | 100.9 | 29.6 | 57.2 |

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.

Sub Experiment main effect

|  |  |  |
| --- | --- | --- |
|  | a | b |
| mean | 68.4 | 57.0 |
| N | 2816 | 2801 |

The two sub experiments have surprisingly different means. This difference is not only statistically significant, it may lead to insight about the differences between the signal sets or the laboratories employed in the two sub-experiments.

Signal Category main effect

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m | s | sbm | som |
| mean | 58.8 | 65.0 | 62.9 | 63.9 |
| N | 1873 | 1880 | 936 | 928 |

This variable is highly statistically significant. Further, the signal categories do have means that do differ somewhat and so there may be some practical difference between the signal categories in this experiment.

Codec by Signal Category (Codec:SigCat) interaction effect

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Codec | SigCat | | | | |
|  |  | m | s | sbm | som |
| Codec1 | mean | 76.9 | 68.9 | 75.3 | 78.8 |
|  | N | 235 | 235 | 117 | 116 |
| Codec2 | mean | 58.7 | 79.2 | 65.6 | 65.9 |
|  | N | 235 | 235 | 117 | 116 |
| Codec3 | mean | 77.2 | 71.5 | 75.9 | 78.7 |
|  | N | 235 | 235 | 117 | 116 |
| AAC | mean | 47.6 | 51.6 | 52.3 | 51.9 |
|  | N | 235 | 235 | 117 | 116 |
| AMR-WB | mean | 37.0 | 57.0 | 46.8 | 48.6 |
|  | N | 235 | 235 | 117 | 116 |
| hidref | mean | 99.8 | 99.9 | 100.1 | 100.0 |
|  | N | 235 | 235 | 117 | 116 |
| lp3500 | mean | 23.7 | 29.9 | 30.1 | 30.7 |
|  | N | 235 | 235 | 117 | 116 |
| lp7000 | mean | 49.6 | 61.7 | 57.0 | 56.7 |
|  | N | 235 | 235 | 117 | 116 |

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 m and s categories is ±1.8, while the width of the 95% confidence intervals for the som and sbm categories is ±2.5.

Signal main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | m\_ot\_x\_2 | m\_ot\_x\_3 | m\_po\_x\_1 | m\_po\_x\_2 | m\_po\_x\_3 | m\_po\_x\_4 |
| mean | 69.4 | 64.7 | 64.1 | 62.5 | 65.5 | 61.6 |
| N | 240 | 232 | 240 | 240 | 224 | 240 |
|  | m\_si\_x\_1 | m\_si\_x\_2 | s\_cl\_2t\_1 | s\_cl\_2t\_2 | s\_cl\_ft\_3 | s\_cl\_mt\_2 |
| mean | 51.5 | 62.0 | 58.7 | 61.9 | 61.7 | 58.7 |
| N | 240 | 224 | 240 | 232 | 232 | 240 |
|  | s\_no\_2t\_1 | s\_no\_2t\_2 | s\_no\_ft\_4 | s\_no\_mt\_1 | sbm\_js\_x\_2 | sbm\_ms\_x\_1 |
| mean | 66.2 | 62.9 | 65.0 | 66.2 | 61.6 | 58.1 |
| N | 232 | 232 | 232 | 240 | 240 | 240 |
|  | sbm\_sm\_x\_2 | sbm\_sm\_x\_5 | som\_nt\_x\_1 | som\_ot\_x\_1 | som\_ot\_x\_2 | som\_ot\_x\_3 |
| mean | 66.7 | 64.5 | 65.8 | 60.1 | 61.2 | 63.8 |
| N | 216 | 240 | 216 | 240 | 240 | 232 |

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CT | DY | FhG | FT |
| mean | 64.4 | 69.7 | 55.7 | 60.8 |
| N | 1440 | 1392 | 1424 | 1368 |

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. However, the effect of this interaction compared to, say, the listener differences, the signal differences or the codec-signal interaction is relatively small. If this interaction had not been included in the statistical model, the residual standard error would have been about 10% larger.

### Post-screening of data

Of the 720 sets of 8 judgments (one for each codec, reference codec, and anchor) in this experiment, 17 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. However, in the analysis of variance, the standard error of the residuals, and thus all confidence interval widths, increases by less than 1%.

## Test A4a and A4b

### Test parameters and systems under test

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Symbol** |
| Experiment | A4a and A4b |  |
| Bit Rate | 24 kbps |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+ | Codec 1 |
| AMR-WB+ | Codec 2 |
| CT | Codec 3 |
| Reference codecs | AAC | AAC |
| AMR-WB, 18.25 kbps, 16 kHz sampling rate | AMR-WB |
| Anchors and references | Open Reference |  |
| Hidden Reference | HR |
| 7.0 kHz Lowpass, 6 dB attenuated side channel | LP7.0-S6 |
| 7.0 kHz Lowpass,12 dB attenuated side channel | LP7.0-S12 |
| 3.5 kHz Lowpass, 12 dB attenuated side channel | LP3.5-S12 |

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

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | Codec3 | AAC | AMR-WB | hidref | lp3500\_s12 | lp7000\_s12 | lp7000\_s6 |
| Average | 55.3 | 61.3 | 67.1 | 34.8 | 44.8 | 99.8 | 31.3 | 57.8 | 61.1 |
| Lower Bound | 52.9 | 59.1 | 64.9 | 32.7 | 42.9 | 99.8 | 29.6 | 56.0 | 59.3 |
| Upper Bound | 57.7 | 63.6 | 69.4 | 36.9 | 46.7 | 99.9 | 33.0 | 59.6 | 62.9 |

The following 3 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 | | | Codec 3 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| m\_ch\_x\_1 | 71.0 | 54.5 | 62.8 | 65.2 | 49.8 | 57.5 | 90.5 | 81.4 | 86.0 |
| m\_cl\_x\_1 | 63.6 | 43.5 | 53.6 | 70.4 | 53.1 | 61.8 | 77.2 | 57.6 | 67.4 |
| m\_cl\_x\_2 | 73.9 | 56.3 | 65.1 | 78.0 | 60.5 | 69.2 | 88.5 | 77.3 | 82.9 |
| m\_ot\_x\_1 | 72.0 | 52.6 | 62.3 | 65.8 | 47.5 | 56.6 | 91.7 | 82.3 | 87.0 |
| m\_ot\_x\_2 | 79.1 | 65.5 | 72.3 | 59.9 | 42.9 | 51.4 | 88.0 | 76.6 | 82.3 |
| m\_ot\_x\_3 | 79.5 | 67.1 | 73.3 | 74.5 | 59.3 | 66.9 | 93.1 | 85.7 | 89.4 |
| m\_po\_x\_1 | 63.0 | 48.8 | 55.9 | 35.0 | 20.5 | 27.8 | 82.3 | 66.4 | 74.3 |
| m\_si\_x\_1 | 35.4 | 18.4 | 26.9 | 45.5 | 29.0 | 37.3 | 65.2 | 46.6 | 55.9 |
| s\_cl\_2t\_1 | 23.0 | 13.3 | 18.2 | 77.1 | 57.9 | 67.5 | 43.8 | 28.4 | 36.1 |
| s\_cl\_2t\_2 | 56.5 | 40.4 | 48.5 | 75.4 | 58.4 | 66.9 | 75.3 | 58.7 | 67.0 |
| s\_cl\_2t\_3 | 48.7 | 27.7 | 38.2 | 80.6 | 66.2 | 73.4 | 56.1 | 37.2 | 46.7 |
| s\_cl\_mt\_1 | 50.1 | 32.4 | 41.2 | 70.8 | 51.2 | 61.0 | 75.0 | 55.7 | 65.4 |
| s\_no\_2t\_1 | 60.0 | 42.0 | 51.0 | 61.3 | 42.3 | 51.8 | 75.5 | 57.2 | 66.4 |
| s\_no\_2t\_2 | 81.4 | 64.5 | 73.0 | 82.3 | 68.0 | 75.2 | 84.1 | 66.2 | 75.2 |
| s\_no\_ft\_1 | 64.1 | 45.9 | 55.0 | 57.9 | 38.3 | 48.1 | 68.6 | 49.9 | 59.3 |
| s\_no\_ft\_2 | 57.1 | 38.7 | 47.9 | 75.7 | 59.4 | 67.6 | 52.9 | 33.4 | 43.1 |
| sbm\_ms\_x\_1 | 51.9 | 35.3 | 43.6 | 75.0 | 59.9 | 67.4 | 48.9 | 33.5 | 41.2 |
| sbm\_sm\_x\_1 | 79.3 | 62.5 | 70.9 | 86.5 | 74.0 | 80.3 | 77.8 | 61.1 | 69.4 |
| sbm\_sm\_x\_2 | 38.6 | 22.6 | 30.6 | 59.6 | 39.8 | 49.7 | 60.6 | 43.6 | 52.1 |
| sbm\_sm\_x\_3 | 63.8 | 45.3 | 54.6 | 88.0 | 78.1 | 83.0 | 75.2 | 58.8 | 67.0 |
| som\_fi\_x\_1 | 85.7 | 72.3 | 79.0 | 64.1 | 44.6 | 54.4 | 91.0 | 80.8 | 85.9 |
| som\_fi\_x\_2 | 74.8 | 53.6 | 64.2 | 72.3 | 54.1 | 63.2 | 79.0 | 60.2 | 69.6 |
| som\_nt\_x\_1 | 73.9 | 55.3 | 64.6 | 60.4 | 44.5 | 52.4 | 86.7 | 75.7 | 81.2 |
| som\_ot\_x\_1 | 68.5 | 49.9 | 59.2 | 76.2 | 59.2 | 67.7 | 77.4 | 58.1 | 67.7 |

### 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 | 2993494 | 374187 | 1493.3 | < 2.2e-16 \*\*\* |
| Sub | 1 | 185315 | 185315 | 739.6 | < 2.2e-16 \*\*\* |
| SigCat | 3 | 17435 | 5812 | 23.2 | 6.07e-15 \*\*\* |
| Signal | 19 | 80881 | 4257 | 17.0 | < 2.2e-16 \*\*\* |
| Site | 2 | 601061 | 300531 | 1199.4 | < 2.2e-16 \*\*\* |
| Subject | 56 | 697285 | 12452 | 49.7 | < 2.2e-16 \*\*\* |
| Codec:Signal | 24 | 277209 | 11550 | 46.1 | < 2.2e-16 \*\*\* |
| Codec:Site | 24 | 181124 | 7547 | 30.1 | < 2.2e-16 \*\*\* |
| Residuals | 8079 | 2024414 | 251 |  |  |

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 | Codec3 | AAC | AMR-WB | hidref | lp3500\_s12 | lp7000\_s12 | lp7000\_s6 |
| mean | 55.3 | 61.3 | 67.1 | 34.8 | 44.8 | 99.8 | 31.3 | 57.8 | 61.1 |
| N | 686 | 686 | 686 | 686 | 686 | 686 | 686 | 686 | 686 |
| Lower Bound | 54.1 | 60.2 | 65.9 | 33.6 | 43.6 | 98.7 | 30.1 | 56.6 | 59.9 |
| Upper Bound | 56.4 | 62.5 | 68.3 | 36.0 | 46.0 | 101.0 | 32.5 | 59.0 | 62.3 |

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.

Sub Experiment main effect

|  |  |  |
| --- | --- | --- |
|  | a | b |
| mean | 52.5 | 62.0 |
| N | 3213 | 2961 |

The two sub experiments have surprisingly different means. This difference is not only statistically significant, it may lead to insight about the differences between the signal sets or the laboratories employed in the two sub-experiments.

Signal Category main effect

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m | s | sbm | som |
| mean | 56.1 | 55.2 | 58.2 | 58.7 |
| N | 2079 | 2052 | 1026 | 1017 |

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 | | | | |
|  |  | m | s | sbm | som |
| Codec1 | mean | 58.7 | 46.2 | 49.3 | 67.0 |
|  | N | 231 | 228 | 114 | 113 |
| Codec2 | mean | 53.1 | 63.7 | 69.6 | 59.1 |
|  | N | 231 | 228 | 114 | 113 |
| Codec3 | mean | 77.8 | 57.3 | 56.9 | 76.4 |
|  | N | 231 | 228 | 114 | 113 |
| AAC | mean | 38.6 | 22.3 | 43.4 | 34.9 |
|  | N | 231 | 228 | 114 | 113 |
| AMR-WB | mean | 34.2 | 51.1 | 46.5 | 47.6 |
|  | N | 231 | 228 | 114 | 113 |
| hidref | mean | 99.8 | 99.9 | 99.9 | 99.8 |
|  | N | 231 | 228 | 114 | 113 |
| lp3500\_s12 |  | 29.7 | 32.7 | 33.7 | 29.1 |
|  |  | 231 | 228 | 114 | 113 |
| lp7000\_s12 | mean | 55.4 | 59.7 | 59.8 | 56.4 |
|  | N | 231 | 228 | 114 | 113 |
| lp7000\_s6 | mean | 57.8 | 63.7 | 64.8 | 58.3 |
|  | N | 231 | 228 | 114 | 113 |

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 m and s categories is ±2.1, while the width of the 95% confidence intervals for the som and sbm categories is ±2.9.

Signal main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | m\_ch\_x\_1 | m\_cl\_x\_1 | m\_cl\_x\_2 | m\_ot\_x\_1 | m\_ot\_x\_2 | m\_ot\_x\_3 |
| mean | 61.4 | 56.2 | 59.0 | 57.2 | 61.5 | 58.7 |
| N | 261 | 270 | 234 | 252 | 261 | 261 |
|  | m\_po\_x\_1 | m\_si\_x\_1 | s\_cl\_2t\_1 | s\_cl\_2t\_2 | s\_cl\_2t\_3 | s\_cl\_mt\_1 |
| mean | 53.2 | 49.9 | 53.8 | 58.1 | 54.6 | 57.6 |
| N | 270 | 270 | 270 | 270 | 225 | 270 |
|  | s\_no\_2t\_1 | s\_no\_2t\_2 | s\_no\_ft\_1 | s\_no\_ft\_2 | sbm\_ms\_x\_1 | sbm\_sm\_x\_1 |
| mean | 58.7 | 60.0 | 53.1 | 60.3 | 55.2 | 60.6 |
| N | 270 | 225 | 252 | 270 | 270 | 243 |
|  | sbm\_sm\_x\_2 | sbm\_sm\_x\_3 | som\_fi\_x\_1 | som\_fi\_x\_2 | som\_nt\_x\_1 | som\_ot\_x\_1 |
| mean | 53.6 | 59.4 | 60.8 | 55.4 | 58.8 | 52.5 |
| N | 270 | 243 | 261 | 252 | 270 | 234 |

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 | Nokia | NTT-AT | T-Sys |
| mean | 68.5 | 60.4 | 53.5 | 45.7 |
| N | 1602 | 1503 | 1458 | 1611 |

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. However, the effect of this interaction compared to, say, the listener differences, the signal differences or the codec-signal interaction is relatively small. If this interaction had not been included in the statistical model, the residual standard error would have been about 9% larger.

### Post-screening of data

Of the 720 sets of 9 judgments (one for each codec, reference codec, and anchor) in this experiment, 34 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. However, in the analysis of variance, the standard error of the residuals, and thus all confidence interval widths, increases by about 1%.

## Test B1a and B1b

### Test parameters and systems under test

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Symbol** |
| Experiment | B1a and B1b |  |
| Bit Rate | 14 kbps |  |
| Signal | Mono, 16 kHz input and output sampling rate |  |
| Candidate codecs | AAC+ | Codec 1 |
| AMR-WB+ | Codec 2 |
| CT | Codec 3 |
| Reference codecs | AAC | AAC |
| AMR-WB, 14.25 kbps, 16 kHz sampling rate | AMR-WB |
| Anchors and references | Open Reference |  |
| Hidden Reference | HR |
| 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.



Only Codec2 out-performs both reference codecs in this experiment. Although Codec1 and Codec3 both out-perform AAC, they perform not statistically different from AMR-WB. The following table shows the numerical values plotted in the chart above.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | Codec3 | AAC | AMR-WB | hidref | lp3500 | lp7000 |
| Average | 45.4 | 50.7 | 44.4 | 30.7 | 46.2 | 100.0 | 31.7 | 65.3 |
| Lower Bound | 43.9 | 49.0 | 42.9 | 29.2 | 44.4 | 100.0 | 30.3 | 63.9 |
| Upper Bound | 47.0 | 52.4 | 46.0 | 32.3 | 47.9 | 100.0 | 33.1 | 66.7 |

The following 3 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 | | | Codec 3 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| m\_ch\_x\_1 | 71.5 | 60.8 | 66.1 | 68.2 | 57.7 | 62.9 | 71.1 | 59.1 | 65.1 |
| m\_cl\_x\_1 | 55.7 | 39.5 | 47.6 | 71.7 | 59.3 | 65.5 | 62.1 | 46.0 | 54.0 |
| m\_po\_x\_2 | 61.8 | 49.0 | 55.4 | 56.5 | 43.2 | 49.9 | 56.1 | 44.6 | 50.3 |
| m\_po\_x\_3 | 66.7 | 54.3 | 60.5 | 51.6 | 39.4 | 45.5 | 63.1 | 51.0 | 57.1 |
| m\_po\_x\_4 | 46.9 | 37.8 | 42.3 | 41.9 | 31.7 | 36.8 | 47.2 | 38.1 | 42.6 |
| m\_po\_x\_5 | 47.4 | 38.8 | 43.1 | 45.4 | 35.6 | 40.5 | 47.9 | 39.6 | 43.7 |
| m\_po\_x\_6 | 54.1 | 42.5 | 48.3 | 35.5 | 26.2 | 30.9 | 48.3 | 41.4 | 44.8 |
| m\_si\_x\_2 | 43.8 | 35.7 | 39.7 | 37.3 | 29.6 | 33.4 | 45.4 | 36.2 | 40.8 |
| s\_cl\_2t\_2 | 44.3 | 34.7 | 39.5 | 50.6 | 41.0 | 45.8 | 44.9 | 35.2 | 40.0 |
| s\_cl\_2t\_4 | 43.3 | 31.9 | 37.6 | 76.1 | 62.3 | 69.2 | 42.1 | 31.0 | 36.6 |
| s\_cl\_ft\_3 | 49.0 | 36.4 | 42.7 | 79.2 | 67.3 | 73.2 | 52.2 | 39.4 | 45.8 |
| s\_cl\_mt\_1 | 35.9 | 27.5 | 31.7 | 51.8 | 42.6 | 47.2 | 35.1 | 27.1 | 31.1 |
| s\_no\_2t\_2 | 69.5 | 56.8 | 63.2 | 64.3 | 49.4 | 56.8 | 65.5 | 50.9 | 58.2 |
| s\_no\_2t\_3 | 37.2 | 29.3 | 33.3 | 58.3 | 47.8 | 53.0 | 37.9 | 29.5 | 33.7 |
| s\_no\_ft\_2 | 66.8 | 54.0 | 60.4 | 66.5 | 51.8 | 59.1 | 62.5 | 50.1 | 56.3 |
| s\_no\_ft\_4 | 34.2 | 26.0 | 30.1 | 47.5 | 39.3 | 43.4 | 35.4 | 27.4 | 31.4 |
| sbm\_js\_x\_2 | 46.1 | 34.6 | 40.3 | 71.6 | 59.8 | 65.7 | 44.2 | 30.6 | 37.4 |
| sbm\_ms\_x\_1 | 46.8 | 34.3 | 40.6 | 67.9 | 53.5 | 60.7 | 47.6 | 35.7 | 41.7 |
| sbm\_sj\_x\_1 | 45.0 | 34.1 | 39.6 | 47.5 | 34.6 | 41.0 | 41.4 | 32.2 | 36.8 |
| sbm\_sm\_x\_4 | 37.0 | 28.2 | 32.6 | 52.4 | 42.9 | 47.6 | 37.6 | 29.3 | 33.4 |
| som\_ad\_x\_1 | 62.4 | 51.6 | 57.0 | 58.1 | 45.7 | 51.9 | 63.7 | 53.7 | 58.7 |
| som\_ot\_x\_2 | 40.4 | 30.6 | 35.5 | 54.5 | 44.5 | 49.5 | 39.8 | 30.1 | 34.9 |
| som\_ot\_x\_4 | 72.5 | 58.4 | 65.5 | 68.1 | 54.4 | 61.2 | 66.7 | 51.9 | 59.3 |
| som\_ot\_x\_6 | 50.5 | 40.3 | 45.4 | 33.2 | 23.9 | 28.5 | 47.8 | 37.8 | 42.8 |

### 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 | 7 | 3303832 | 471976 | 3463.3 | < 2.2e-16 \*\*\* |
| Sub | 1 | 282871 | 282871 | 2075.7 | < 2.2e-16 \*\*\* |
| SigCat | 3 | 13911 | 4637 | 34.0 | < 2.2e-16 \*\*\* |
| Signal | 19 | 32501 | 1711 | 12.6 | < 2.2e-16 \*\*\* |
| Site | 2 | 233235 | 116617 | 855.7 | < 2.2e-16 \*\*\* |
| Subject | 56 | 404146 | 7217 | 53.0 | < 2.2e-16 \*\*\* |
| Codec:Signal | 21 | 181389 | 8638 | 63.4 | < 2.2e-16 \*\*\* |
| Codec:Site | 21 | 96737 | 4607 | 33.8 | < 2.2e-16 \*\*\* |
| Residuals | 7461 | 1016783 | 136 |  |  |

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 | Codec3 | AAC | AMR-WB | hidref | lp3500 | lp7000 |
| mean | 45.5 | 50.7 | 44.4 | 30.7 | 46.2 | 100.0 | 31.7 | 65.3 |
| N | 713 | 713 | 713 | 713 | 713 | 713 | 713 | 713 |
| Lower Bound | 44.6 | 49.9 | 43.6 | 29.9 | 45.2 | 99.1 | 30.8 | 64.6 |
| Upper Bound | 46.3 | 51.6 | 45.3 | 31.6 | 47.0 | 100.9 | 32.5 | 66.2 |

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.

Sub Experiment main effect

|  |  |  |
| --- | --- | --- |
|  | a | b |
| mean | 57.8 | 45.8 |
| N | 2824 | 2880 |

The two sub experiments have surprisingly different means. This difference is not only statistically significant, it may lead to insight about the differences between the signal sets or the laboratories employed in the two sub-experiments.

Signal Category main effect

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m | s | sbm | som |
| mean | 50.5 | 53.3 | 50.5 | 53.1 |
| N | 1920 | 1896 | 944 | 944 |

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 | | | | |
|  |  | m | s | sbm | som |
| Codec1 | mean | 50.4 | 42.3 | 38.2 | 50.8 |
|  | N | 240 | 237 | 118 | 118 |
| Codec2 | mean | 45.7 | 55.9 | 53.8 | 47.6 |
|  | N | 240 | 237 | 118 | 118 |
| Codec3 | mean | 49.8 | 41.6 | 37.3 | 48.9 |
|  | N | 240 | 237 | 118 | 118 |
| AAC | mean | 38.0 | 28.6 | 26.5 | 29.7 |
|  | N | 240 | 237 | 118 | 118 |
| AMR-WB | mean | 29.7 | 57.5 | 48.5 | 49.3 |
|  | N | 240 | 237 | 118 | 118 |
| hidref | mean | 99.9 | 100.0 | 100.0 | 100.0 |
|  | N | 240 | 237 | 118 | 118 |
| lp3500 | mean | 28.2 | 33.1 | 33.5 | 32.0 |
|  | N | 240 | 237 | 118 | 118 |
| lp7000 | mean | 62.0 | 67.0 | 66.1 | 66.2 |
|  | N | 240 | 237 | 118 | 118 |

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 m and s categories is ±1.5, while the width of the 95% confidence intervals for the som and sbm categories is ±2.1.

Signal main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | m\_ch\_x\_1 | m\_cl\_x\_1 | m\_po\_x\_2 | m\_po\_x\_3 | m\_po\_x\_4 | m\_po\_x\_5 |
| mean | 55.9 | 53.9 | 46.3 | 50.8 | 52.4 | 53.0 |
| N | 240 | 240 | 240 | 240 | 240 | 240 |
|  | m\_po\_x\_6 | m\_si\_x\_2 | s\_cl\_2t\_2 | s\_cl\_2t\_4 | s\_cl\_ft\_3 | s\_cl\_mt\_1 |
| mean | 52.6 | 49.5 | 52.3 | 50.4 | 52.2 | 49.1 |
| N | 240 | 240 | 240 | 232 | 232 | 240 |
|  | s\_no\_2t\_2 | s\_no\_2t\_3 | s\_no\_ft\_2 | s\_no\_ft\_4 | sbm\_js\_x\_2 | sbm\_ms\_x\_1 |
| mean | 52.8 | 53.4 | 53.7 | 50.7 | 51.1 | 50.3 |
| N | 232 | 240 | 240 | 240 | 232 | 240 |
|  | sbm\_sj\_x\_1 | sbm\_sm\_x\_4 | som\_ad\_x\_1 | som\_ot\_x\_2 | som\_ot\_x\_4 | som\_ot\_x\_6 |
| mean | 53.3 | 52.6 | 49.8 | 52.7 | 55.5 | 49.5 |
| N | 232 | 240 | 240 | 232 | 232 | 240 |

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CT | DY | FhG | FT |
| mean | 52.2 | 59.8 | 44.1 | 51.4 |
| N | 1440 | 1416 | 1440 | 1424 |

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. However, the effect of this interaction compared to, say, the listener differences, the signal differences or the codec-signal interaction is relatively small. If this interaction had not been included in the statistical model, the residual standard error would have been about 10% larger.

### Post-screening of data

Of the 720 sets of 8 judgments (one for each codec, reference codec, and anchor) in this experiment, 7 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. However, in the analysis of variance, the standard error of the residuals, and thus all confidence interval widths, increases by less than 1%.

## Test B2a and B2b

### Test parameters and systems under test

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Symbol** |
| Experiment | B2a and B2b |  |
| Bit Rate | 18 kbps |  |
| Signal | Stereo |  |
| Candidate codecs | AAC+ | Codec 1 |
| AMR-WB+ | Codec 2 |
| CT | Codec 3 |
| Reference codecs | AAC | AAC |
| AMR-WB, 18.25 kbps, 16 kHz sampling rate | AMR-WB |
| Anchors and references | Open Reference |  |
| Hidden Reference | HR |
| 7.0 kHz Lowpass, 6 dB attenuated side channel | LP7.0-S6 |
| 7.0 kHz Lowpass,12 dB attenuated side channel | LP7.0-S12 |
| 3.5 kHz Lowpass, 12 dB attenuated side channel | LP3.5-S12 |

### 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.



While all of the candidates codecs out-perform the AAC reference codec, Codec1 fails to outperform AMR-WB, and a more sensitive analysis (section 8.6.3) is needed to determine that Codec2 does indeed out-perform AMR-WB. The following table shows the numerical values plotted in the chart above.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | Codec3 | AAC | AMR-WB | hidref | lp3500\_s12 | lp7000\_s12 | lp7000\_s6 |
| Average | 43.3 | 50.7 | 55.7 | 22.8 | 46.8 | 99.9 | 33.1 | 62.6 | 64.2 |
| Lower Bound | 41.1 | 48.4 | 53.3 | 21.2 | 44.8 | 99.9 | 31.3 | 60.7 | 62.4 |
| Upper Bound | 45.5 | 52.9 | 58.1 | 24.5 | 48.8 | 100.0 | 34.8 | 64.4 | 66.1 |

The following 3 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 | | | Codec 3 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| m\_ch\_x\_1 | 65.0 | 49.6 | 57.3 | 75.9 | 61.0 | 68.5 | 89.1 | 78.2 | 83.7 |
| m\_cl\_x\_1 | 47.6 | 28.6 | 38.1 | 68.2 | 46.6 | 57.4 | 60.2 | 37.8 | 49.0 |
| m\_cl\_x\_2 | 59.3 | 40.7 | 50.0 | 29.1 | 18.9 | 24.0 | 69.1 | 51.3 | 60.2 |
| m\_ot\_x\_1 | 56.4 | 40.8 | 48.6 | 54.2 | 38.2 | 46.2 | 64.0 | 46.7 | 55.3 |
| m\_ot\_x\_8 | 50.6 | 35.2 | 42.9 | 49.5 | 34.7 | 42.1 | 73.7 | 54.7 | 64.2 |
| m\_ot\_x\_9 | 66.5 | 52.7 | 59.6 | 75.2 | 57.8 | 66.5 | 87.5 | 76.1 | 81.8 |
| m\_ot\_x\_a | 58.6 | 41.0 | 49.8 | 30.0 | 19.4 | 24.7 | 79.6 | 64.4 | 72.0 |
| m\_ot\_x\_b | 64.5 | 49.4 | 57.0 | 66.5 | 49.6 | 58.0 | 85.3 | 71.8 | 78.6 |
| s\_cl\_2t\_3 | 38.0 | 22.6 | 30.3 | 67.5 | 51.3 | 59.4 | 49.7 | 31.8 | 40.7 |
| s\_cl\_2t\_4 | 24.3 | 15.0 | 19.6 | 72.5 | 54.4 | 63.5 | 39.7 | 24.6 | 32.2 |
| s\_cl\_2t\_5 | 49.0 | 33.1 | 41.0 | 52.6 | 38.8 | 45.7 | 69.1 | 50.8 | 60.0 |
| s\_cl\_mt\_1 | 45.3 | 29.4 | 37.4 | 69.2 | 51.2 | 60.2 | 72.0 | 54.7 | 63.4 |
| s\_no\_2t\_3 | 34.0 | 20.4 | 27.2 | 59.2 | 42.0 | 50.6 | 41.5 | 25.6 | 33.6 |
| s\_no\_ft\_1 | 49.2 | 32.7 | 40.9 | 51.9 | 33.6 | 42.7 | 55.7 | 37.3 | 46.5 |
| s\_no\_ft\_2 | 47.2 | 30.3 | 38.8 | 60.7 | 43.0 | 51.9 | 45.5 | 28.6 | 37.1 |
| s\_no\_ft\_3 | 64.5 | 43.8 | 54.2 | 55.7 | 33.8 | 44.8 | 53.5 | 30.9 | 42.2 |
| sbm\_js\_x\_1 | 23.1 | 14.8 | 18.9 | 42.5 | 26.3 | 34.4 | 32.1 | 20.2 | 26.2 |
| sbm\_sm\_x\_1 | 57.9 | 38.2 | 48.0 | 76.5 | 62.0 | 69.3 | 74.9 | 55.9 | 65.4 |
| sbm\_sm\_x\_3 | 48.3 | 31.7 | 40.0 | 78.4 | 62.8 | 70.6 | 54.4 | 38.6 | 46.5 |
| sbm\_sm\_x\_4 | 26.4 | 16.2 | 21.3 | 58.4 | 41.9 | 50.2 | 52.1 | 36.1 | 44.1 |
| som\_fi\_x\_1 | 64.9 | 45.9 | 55.4 | 54.7 | 37.6 | 46.1 | 77.7 | 57.2 | 67.5 |
| som\_fi\_x\_2 | 66.9 | 50.7 | 58.8 | 61.5 | 42.1 | 51.8 | 74.7 | 56.9 | 65.8 |
| som\_ot\_x\_5 | 57.9 | 42.1 | 50.0 | 58.2 | 42.9 | 50.6 | 77.5 | 62.2 | 69.9 |
| som\_ot\_x\_6 | 66.3 | 48.5 | 57.4 | 47.3 | 27.4 | 37.3 | 67.0 | 45.6 | 56.3 |

### 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 | 8 | 3584309 | 448039 | 1942.8 | < 2.2e-16 \*\*\* |
| Sub | 1 | 169861 | 169861 | 736.6 | < 2.2e-16 \*\*\* |
| SigCat | 3 | 8504 | 2835 | 12.3 | 5.09e-08 \*\*\* |
| Signal | 19 | 42875 | 2257 | 9.8 | < 2.2e-16 \*\*\* |
| Site | 2 | 622804 | 311402 | 1350.3 | < 2.2e-16 \*\*\* |
| Subject | 56 | 810238 | 14469 | 62.7 | < 2.2e-16 \*\*\* |
| Codec:Signal | 24 | 347422 | 14476 | 62.8 | < 2.2e-16 \*\*\* |
| Codec:Site | 24 | 134199 | 5592 | 24.2 | < 2.2e-16 \*\*\* |
| Residuals | 8196 | 1890101 | 231 |  |  |

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 | Codec3 | AAC | AMR-WB | hidref | lp3500\_s12 | lp7000\_s12 | lp7000\_s6 |
| mean | 43.3 | 50.7 | 55.7 | 22.9 | 46.8 | 99.9 | 33.1 | 62.6 | 64.2 |
| N | 696 | 696 | 696 | 696 | 696 | 696 | 696 | 696 | 696 |
| Lower Bound | 42.2 | 49.5 | 54.5 | 21.7 | 45.7 | 98.8 | 32.0 | 61.4 | 63.1 |
| Upper Bound | 44.4 | 51.8 | 56.8 | 24.0 | 47.9 | 101.1 | 34.2 | 63.7 | 65.4 |

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. *In fact, as shown here, the confidence intervals for Codec2 and AMR-WB do not overlap.*

Sub Experiment main effect

|  |  |  |
| --- | --- | --- |
|  | a | b |
| mean | 48.9 | 57.9 |
| N | 3231 | 3033 |

The two sub experiments have surprisingly different means. This difference is not only statistically significant, it may lead to insight about the differences between the signal sets or the laboratories employed in the two sub-experiments.

Signal Category main effect

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m | s | sbm | som |
| mean | 53.3 | 51.7 | 53.4 | 54.6 |
| N | 2097 | 2097 | 1035 | 1035 |

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 | | | | |
|  |  | m | s | sbm | som |
| Codec1 | mean | 50.4 | 36.0 | 31.6 | 55.3 |
|  | N | 233 | 233 | 115 | 115 |
| Codec2 | mean | 48.1 | 52.4 | 55.6 | 46.6 |
|  | N | 233 | 233 | 115 | 115 |
| Codec3 | mean | 68.0 | 44.4 | 45.2 | 65.0 |
|  | N | 233 | 233 | 115 | 115 |
| AAC | mean | 30.6 | 15.2 | 20.7 | 24.9 |
|  | N | 233 | 233 | 115 | 115 |
| AMR-WB | mean | 31.5 | 53.0 | 59.0 | 43.8 |
|  | N | 233 | 233 | 115 | 115 |
| hidref | mean | 99.9 | 99.9 | 100.0 | 99.9 |
|  | N | 233 | 233 | 115 | 115 |
| lp3500\_s12 | mean | 31.3 | 34.8 | 33.9 | 32.3 |
|  | N | 233 | 233 | 115 | 115 |
| lp7000\_s12 | mean | 58.8 | 64.0 | 66.2 | 61.3 |
|  | N | 233 | 233 | 115 | 115 |
| lp7000\_s6 | mean | 61.1 | 65.9 | 68.0 | 61.9 |
|  | N | 233 | 233 | 115 | 115 |

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 m and s categories is ±2.0, while the width of the 95% confidence intervals for the som and sbm categories is ±2.8.

Signal main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | m\_ch\_x\_1 | m\_cl\_x\_1 | m\_cl\_x\_2 | m\_ot\_x\_1 | m\_ot\_x\_8 | m\_ot\_x\_9 |
| mean | 54.3 | 48.3 | 53.0 | 52.9 | 54.2 | 57.1 |
| N | 261 | 252 | 270 | 270 | 270 | 261 |
|  | m\_ot\_x\_a | m\_ot\_x\_b | s\_cl\_2t\_3 | s\_cl\_2t\_4 | s\_cl\_2t\_5 | s\_cl\_mt\_1 |
| mean | 53.9 | 51.8 | 53.6 | 52.8 | 57.3 | 52.6 |
| N | 270 | 243 | 252 | 270 | 270 | 261 |
|  | s\_no\_2t\_3 | s\_no\_ft\_1 | s\_no\_ft\_2 | s\_no\_ft\_3 | sbm\_js\_x\_1 | sbm\_sm\_x\_1 |
| mean | 53.8 | 49.0 | 55.2 | 51.2 | 50.4 | 56.0 |
| N | 270 | 261 | 270 | 243 | 540 | 504 |
|  | sbm\_sm\_x\_3 | sbm\_sm\_x\_4 | som\_fi\_x\_1 | som\_fi\_x\_2 | som\_ot\_x\_5 | som\_ot\_x\_6 |
| mean | 56.0 | 51.1 | 54.6 | 54.5 | 53.2 | 50.4 |
| N | 486 | 540 | 522 | 522 | 540 | 486 |

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 | Nokia | NTT-AT | T-Sys |
| mean | 64.8 | 56.5 | 49.7 | 41.6 |
| N | 1620 | 1575 | 1458 | 1611 |

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. However, the effect of this interaction compared to, say, the listener differences, the signal differences or the codec-signal interaction is relatively small. If this interaction had not been included in the statistical model, the residual standard error would have been about 7% larger.

### Post-screening of data

Of the 720 sets of 9 judgments (one for each codec, reference codec, and anchor) in this experiment, 24 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. However, in the analysis of variance, the standard error of the residuals, and thus all confidence interval widths, increases by about 1%.

## Test B3a and B3b

### Test parameters and systems under test

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Symbol** |
| Experiment | B3a and B3b |  |
| Bit Rate | 14 kbps |  |
| Signal | Mono |  |
| Channel Error Condition | 3% FER |  |
| Candidate codecs | AAC+ | Codec 1 |
| AMR-WB+ | Codec 2 |
| CT | Codec 3 |
| Reference codecs | AAC | AAC |
| AMR-WB, 14.25 kbps, 16 kHz sampling rate | AMR-WB |
| Anchors and references | Open Reference |  |
| Hidden Reference | HR |
| 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 candidates codecs out-performs both of the reference codecs. The following table shows the numerical values plotted in the chart above.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | Codec3 | AAC | AMR-WB | hidref | lp3500 | lp7000 |
| Average | 43.1 | 52.5 | 44.3 | 32.1 | 24.7 | 100.0 | 37.0 | 64.4 |
| Lower Bound | 41.4 | 50.8 | 42.7 | 30.7 | 23.2 | 100.0 | 35.3 | 62.9 |
| Upper Bound | 44.8 | 54.2 | 46.0 | 33.6 | 26.2 | 100.0 | 38.6 | 65.9 |

The following 3 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 | | | Codec 3 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| m\_ch\_x\_1 | 53.5 | 39.3 | 46.4 | 56.8 | 43.2 | 50.0 | 56.0 | 43.1 | 49.6 |
| m\_cl\_x\_1 | 37.7 | 27.2 | 32.4 | 68.4 | 52.6 | 60.5 | 42.9 | 31.7 | 37.3 |
| m\_cl\_x\_2 | 55.3 | 37.9 | 46.6 | 59.4 | 45.2 | 52.3 | 57.9 | 42.6 | 50.3 |
| m\_ot\_x\_1 | 61.5 | 48.0 | 54.8 | 55.2 | 40.9 | 48.0 | 62.7 | 49.0 | 55.8 |
| m\_ot\_x\_4 | 49.5 | 37.3 | 43.4 | 55.1 | 42.8 | 48.9 | 49.9 | 37.1 | 43.5 |
| m\_ot\_x\_5 | 72.0 | 56.4 | 64.2 | 67.9 | 52.7 | 60.3 | 69.7 | 55.2 | 62.4 |
| m\_ot\_x\_6 | 54.3 | 39.1 | 46.7 | 47.5 | 32.8 | 40.1 | 48.9 | 35.2 | 42.0 |
| m\_ot\_x\_7 | 53.2 | 34.7 | 43.9 | 67.7 | 53.4 | 60.5 | 56.4 | 39.2 | 47.8 |
| s\_cl\_2t\_1 | 43.3 | 31.7 | 37.5 | 72.1 | 58.5 | 65.3 | 43.4 | 32.0 | 37.7 |
| s\_cl\_2t\_2 | 49.1 | 37.3 | 43.2 | 51.8 | 41.7 | 46.7 | 54.4 | 41.1 | 47.8 |
| s\_cl\_2t\_3 | 43.7 | 32.8 | 38.3 | 63.4 | 50.8 | 57.1 | 43.7 | 32.5 | 38.1 |
| s\_cl\_mt\_1 | 40.2 | 28.8 | 34.5 | 49.9 | 36.6 | 43.2 | 39.0 | 27.6 | 33.3 |
| s\_no\_2t\_3 | 44.4 | 31.5 | 37.9 | 70.9 | 58.2 | 64.6 | 44.4 | 31.8 | 38.1 |
| s\_no\_ft\_3 | 55.0 | 38.4 | 46.7 | 52.0 | 35.6 | 43.8 | 52.5 | 38.3 | 45.4 |
| s\_no\_ft\_4 | 49.2 | 35.8 | 42.5 | 74.2 | 55.7 | 64.9 | 48.3 | 35.8 | 42.0 |
| s\_no\_mt\_1 | 39.8 | 28.8 | 34.3 | 52.8 | 40.1 | 46.5 | 41.1 | 28.7 | 34.9 |
| sbm\_js\_x\_1 | 28.1 | 18.4 | 23.2 | 57.3 | 43.0 | 50.1 | 31.2 | 22.6 | 26.9 |
| sbm\_js\_x\_2 | 43.9 | 33.5 | 38.7 | 69.8 | 59.0 | 64.4 | 45.0 | 34.2 | 39.6 |
| sbm\_ms\_x\_1 | 43.4 | 32.5 | 37.9 | 59.7 | 46.0 | 52.9 | 46.6 | 35.5 | 41.0 |
| sbm\_sm\_x\_1 | 55.5 | 40.7 | 48.1 | 52.4 | 38.2 | 45.3 | 58.2 | 44.7 | 51.4 |
| som\_ad\_x\_1 | 50.2 | 37.5 | 43.9 | 57.2 | 43.4 | 50.3 | 49.9 | 38.1 | 44.0 |
| som\_fi\_x\_1 | 59.6 | 45.6 | 52.6 | 60.4 | 46.4 | 53.4 | 61.5 | 47.1 | 54.3 |
| som\_fi\_x\_2 | 58.2 | 44.0 | 51.1 | 60.0 | 47.2 | 53.6 | 58.5 | 45.2 | 51.9 |
| som\_fi\_x\_3 | 55.2 | 41.2 | 48.2 | 50.8 | 37.4 | 44.1 | 54.6 | 41.2 | 47.9 |

### 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 | 7 | 3681304 | 525901 | 3009.3 | < 2.2e-16 \*\*\* |
| Sub | 1 | 101718 | 101718 | 582.0 | < 2.2e-16 \*\*\* |
| SigCat | 3 | 7139 | 2380 | 13.6 | 7.43e-09 \*\*\* |
| Signal | 19 | 22274 | 1172 | 6.7 | < 2.2e-16 \*\*\* |
| Site | 2 | 22730 | 11365 | 65.0 | < 2.2e-16 \*\*\* |
| Subject | 56 | 526486 | 9402 | 53.8 | < 2.2e-16 \*\*\* |
| Codec:Signal | 21 | 136860 | 6517 | 37.3 | < 2.2e-16 \*\*\* |
| Codec:Site | 21 | 173159 | 8246 | 47.2 | < 2.2e-16 \*\*\* |
| Residuals | 7381 | 1289904 | 175 |  |  |

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 | Codec3 | AAC | AMR-WB | hidref | lp3500 | lp7000 |
| mean | 43.1 | 52.5 | 44.4 | 32.1 | 24.7 | 100.0 | 37.0 | 64.4 |
| N | 702 | 702 | 702 | 702 | 702 | 702 | 702 | 702 |
| Lower Bound | 42.1 | 51.5 | 43.4 | 31.1 | 23.8 | 99.0 | 36.0 | 63.4 |
| Upper Bound | 44.1 | 53.5 | 45.3 | 33.1 | 25.7 | 101.0 | 37.9 | 65.4 |

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.

Sub Experiment main effect

|  |  |  |
| --- | --- | --- |
|  | a | b |
| mean | 53.5 | 46.1 |
| N | 2768 | 2848 |

The two sub experiments have surprisingly different means. This difference is not only statistically significant, it may lead to insight about the differences between the signal sets or the laboratories employed in the two sub-experiments.

Signal Category main effect

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m | s | sbm | som |
| mean | 50.9 | 49.8 | 48.2 | 50.2 |
| N | 1824 | 1896 | 944 | 952 |

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 | | | | |
|  |  | m | s | sbm | som |
| Codec1 | mean | 47.2 | 39.3 | 37.1 | 49.0 |
|  | N | 228 | 237 | 118 | 119 |
| Codec2 | mean | 52.5 | 53.9 | 53.2 | 50.3 |
|  | N | 228 | 237 | 118 | 119 |
| Codec3 | mean | 48.5 | 39.6 | 39.8 | 49.6 |
|  | N | 228 | 237 | 118 | 119 |
| AAC | mean | 45.5 | 26.0 | 27.5 | 30.0 |
|  | N | 228 | 237 | 118 | 119 |
| AMR-WB | mean | 16.1 | 32.2 | 24.4 | 26.0 |
|  | N | 228 | 237 | 118 | 119 |
| hidref | mean | 100.1 | 100.0 | 100.0 | 100.0 |
|  | N | 228 | 237 | 118 | 119 |
| lp3500 | mean | 35.5 | 38.9 | 38.8 | 34.6 |
|  | N | 228 | 237 | 118 | 119 |
| lp7000 | mean | 61.6 | 68.6 | 65.1 | 62.1 |
|  | N | 228 | 237 | 118 | 119 |

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 m and s categories is ±1.7, while the width of the 95% confidence intervals for the som and sbm categories is ±2.4.

Signal main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | m\_ch\_x\_1 | m\_cl\_x\_1 | m\_cl\_x\_2 | m\_ot\_x\_1 | m\_ot\_x\_4 | m\_ot\_x\_5 |
| mean | 48.3 | 48.8 | 52.5 | 48.9 | 49.3 | 52.9 |
| N | 232 | 232 | 240 | 224 | 240 | 216 |
|  | m\_ot\_x\_6 | m\_ot\_x\_7 | s\_cl\_2t\_1 | s\_cl\_2t\_2 | s\_cl\_2t\_3 | s\_cl\_mt\_1 |
| mean | 47.5 | 50.3 | 50.9 | 51.2 | 49.0 | 46.0 |
| N | 232 | 208 | 240 | 232 | 240 | 240 |
|  | s\_no\_2t\_3 | s\_no\_ft\_3 | s\_no\_ft\_4 | s\_no\_mt\_1 | sbm\_js\_x\_1 | sbm\_js\_x\_2 |
| mean | 51.1 | 50.0 | 50.9 | 49.1 | 47.2 | 50.7 |
| N | 232 | 240 | 232 | 240 | 232 | 240 |
|  | sbm\_ms\_x\_1 | sbm\_sm\_x\_1 | som\_ad\_x\_1 | som\_fi\_x\_1 | som\_fi\_x\_2 | som\_fi\_x\_3 |
| mean | 48.6 | 52.4 | 47.5 | 51.2 | 50.0 | 50.3 |
| N | 232 | 240 | 232 | 240 | 240 | 240 |

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CT | DY | FhG | FT |
| mean | 50.1 | 52.3 | 47.4 | 49.4 |
| N | 1424 | 1328 | 1440 | 1424 |

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. However, the effect of this interaction compared to, say, the listener differences, the signal differences or the codec-signal interaction is relatively small. If this interaction had not been included in the statistical model, the residual standard error would have been about 13% larger.

### Post-screening of data

Of the 720 sets of 8 judgments (one for each codec, reference codec, and anchor) in this experiment, 18 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. However, in the analysis of variance, the standard error of the residuals, and thus all confidence interval widths, increases by about 2%.

## Test B4a and B4b

### Test parameters and systems under test

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Symbol** |
| Experiment | B4a and B4b |  |
| Bit Rate | 24 kbps |  |
| Signal | Stereo |  |
| Channel Error Condition | 3% FER |  |
| Candidate codecs | AAC+ | Codec 1 |
| AMR-WB+ | Codec 2 |
| CT | Codec 3 |
| Reference codecs | AAC | AAC |
| AMR-WB, 23.85 kbps, 16 kHz sampling rate, mono | AMR-WB |
| Anchors and references | Open Reference |  |
| Hidden Reference | HR |
| 7.0 kHz Lowpass, 6 dB attenuated side channel | LP7.0-S6 |
| 7.0 kHz Lowpass,12 dB attenuated side channel | LP7.0-S12 |
| 3.5 kHz Lowpass, 12 dB attenuated side channel | LP3.5-S12 |

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

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Codec1 | Codec2 | Codec3 | AAC | AMR-WB | hidref | lp3500\_s12 | lp7000\_s12 | lp7000\_s6 |
| Average | 48.9 | 53.3 | 58.0 | 33.1 | 22.0 | 100.0 | 32.4 | 60.7 | 62.9 |
| Lower Bound | 46.6 | 51.1 | 55.6 | 31.2 | 20.5 | 99.9 | 30.8 | 58.9 | 61.1 |
| Upper Bound | 51.2 | 55.4 | 60.4 | 35.1 | 23.6 | 100.0 | 34.0 | 62.5 | 64.7 |

The following 3 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 | | | Codec 3 | | |
|  | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean | Upper Bound | Lower Bound | Mean |
| m\_ot\_x\_4 | 65.2 | 47.8 | 56.5 | 58.1 | 44.2 | 51.1 | 80.6 | 61.4 | 71.0 |
| m\_ot\_x\_5 | 79.4 | 64.3 | 71.9 | 65.3 | 45.3 | 55.3 | 93.0 | 80.2 | 86.6 |
| m\_ot\_x\_8 | 58.2 | 38.5 | 48.3 | 55.7 | 40.4 | 48.0 | 73.2 | 53.6 | 63.4 |
| m\_ot\_x\_9 | 70.3 | 50.3 | 60.3 | 67.2 | 47.6 | 57.4 | 84.6 | 65.0 | 74.8 |
| m\_po\_x\_2 | 56.1 | 40.4 | 48.3 | 51.1 | 37.1 | 44.1 | 72.6 | 54.1 | 63.3 |
| m\_po\_x\_3 | 67.4 | 50.4 | 58.9 | 65.9 | 49.5 | 57.7 | 82.0 | 67.6 | 74.8 |
| m\_po\_x\_7 | 54.0 | 38.8 | 46.4 | 54.1 | 38.2 | 46.1 | 80.4 | 63.3 | 71.9 |
| m\_si\_x\_3 | 72.2 | 51.5 | 61.9 | 70.6 | 47.8 | 59.2 | 84.4 | 64.9 | 74.7 |
| s\_cl\_2t\_1 | 30.4 | 19.3 | 24.9 | 61.8 | 42.1 | 52.0 | 37.9 | 25.4 | 31.6 |
| s\_cl\_2t\_3 | 44.7 | 27.5 | 36.1 | 70.7 | 51.0 | 60.9 | 48.7 | 29.6 | 39.1 |
| s\_cl\_2t\_5 | 55.1 | 38.4 | 46.7 | 66.9 | 47.8 | 57.4 | 69.7 | 50.4 | 60.0 |
| s\_cl\_mt\_2 | 32.3 | 18.4 | 25.3 | 63.3 | 44.1 | 53.7 | 40.5 | 25.2 | 32.8 |
| s\_no\_2t\_1 | 59.8 | 41.9 | 50.9 | 60.2 | 42.1 | 51.2 | 55.5 | 38.1 | 46.8 |
| s\_no\_ft\_1 | 41.6 | 27.3 | 34.4 | 52.5 | 34.5 | 43.5 | 53.4 | 36.2 | 44.8 |
| s\_no\_ft\_3 | 55.1 | 36.5 | 45.8 | 54.2 | 36.4 | 45.3 | 45.7 | 27.9 | 36.8 |
| s\_no\_mt\_1 | 40.5 | 24.2 | 32.3 | 62.2 | 41.7 | 51.9 | 52.9 | 33.0 | 43.0 |
| sbm\_js\_x\_2 | 52.0 | 35.6 | 43.8 | 52.4 | 37.8 | 45.1 | 58.4 | 41.1 | 49.8 |
| sbm\_sm\_x\_1 | 66.6 | 48.6 | 57.6 | 76.1 | 60.1 | 68.1 | 67.9 | 50.1 | 59.0 |
| sbm\_sm\_x\_2 | 34.3 | 21.6 | 28.0 | 57.9 | 41.0 | 49.5 | 49.5 | 30.6 | 40.1 |
| sbm\_sm\_x\_6 | 68.6 | 44.7 | 56.6 | 78.8 | 62.1 | 70.5 | 66.7 | 42.3 | 54.5 |
| som\_fi\_x\_3 | 67.3 | 49.3 | 58.3 | 57.3 | 41.0 | 49.1 | 75.2 | 58.9 | 67.0 |
| som\_fi\_x\_4 | 50.8 | 33.8 | 42.3 | 57.4 | 40.2 | 48.8 | 63.1 | 46.0 | 54.6 |
| som\_ot\_x\_3 | 70.0 | 52.3 | 61.2 | 67.3 | 51.2 | 59.2 | 76.1 | 58.2 | 67.1 |
| som\_ot\_x\_5 | 70.5 | 52.3 | 61.4 | 55.4 | 39.4 | 47.4 | 85.9 | 71.7 | 78.8 |

### 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 | 3887175 | 485897 | 1959.7 | < 2.2e-16 \*\*\* |
| Sub | 1 | 144206 | 144206 | 581.6 | < 2.2e-16 \*\*\* |
| SigCat | 3 | 15192 | 5064 | 20.4 | 3.51e-13 \*\*\* |
| Signal | 19 | 48865 | 2572 | 10.4 | < 2.2e-16 \*\*\* |
| Site | 2 | 603654 | 301827 | 1217.3 | < 2.2e-16 \*\*\* |
| Subject | 56 | 576837 | 10301 | 41.5 | < 2.2e-16 \*\*\* |
| Codec:Signal | 24 | 300923 | 12538 | 50.6 | < 2.2e-16 \*\*\* |
| Codec:Site | 24 | 162969 | 6790 | 27.4 | < 2.2e-16 \*\*\* |
| Residuals | 8232 | 2041092 | 248 |  |  |

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 | Codec3 | AAC | AMR-WB | hidref | lp3500\_s12 | lp7000\_s12 | lp7000\_s6 |
| mean | 48.9 | 53.3 | 58.0 | 33.1 | 22.0 | 100.0 | 32.4 | 60.7 | 62.9 |
| N | 930 | 930 | 930 | 930 | 930 | 930 | 930 | 930 | 930 |
| Lower Bound | 47.7 | 52.1 | 56.8 | 32.0 | 20.9 | 98.8 | 31.2 | 59.5 | 61.7 |
| Upper Bound | 50.1 | 54.4 | 59.2 | 34.3 | 23.2 | 101.1 | 33.5 | 61.9 | 64.1 |

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.

Sub Experiment main effect

|  |  |  |
| --- | --- | --- |
|  | a | b |
| mean | 48.3 | 56.6 |
| N | 3231 | 3051 |

The two sub experiments have surprisingly different means. This difference is not only statistically significant, it may lead to insight about the differences between the signal sets or the laboratories employed in the two sub-experiments.

Signal Category main effect

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m | s | sbm | som |
| mean | 53.5 | 50.2 | 52.4 | 53.4 |
| N | 2088 | 2106 | 1026 | 1062 |

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 | | | | |
|  |  | m | s | sbm | som |
| Codec1 | mean | 56.4 | 37.1 | 46.0 | 56.0 |
| rep | N | 232 | 234 | 228 | 236 |
| Codec2 | mean | 52.2 | 52.0 | 57.8 | 51.1 |
| rep | N | 232 | 234 | 228 | 236 |
| Codec3 | mean | 72.4 | 41.8 | 50.6 | 67.0 |
| rep | N | 232 | 234 | 228 | 236 |
| AAC | mean | 41.1 | 21.2 | 38.8 | 31.7 |
| rep | N | 232 | 234 | 228 | 236 |
| AMR-WB | mean | 16.6 | 29.7 | 21.8 | 20.0 |
| rep | N | 232 | 234 | 228 | 236 |
| hidref | mean | 99.8 | 100.0 | 100.1 | 99.9 |
| rep | N | 232 | 234 | 228 | 236 |
| lp3500\_s12 | mean | 30.3 | 36.4 | 30.9 | 31.9 |
| rep | N | 232 | 234 | 228 | 236 |
| lp7000\_s12 | mean | 55.0 | 65.3 | 61.7 | 60.7 |
| rep | N | 232 | 234 | 228 | 236 |
| lp7000\_s6 | mean | 57.4 | 67.8 | 64.0 | 62.4 |
| rep | N | 232 | 234 | 228 | 236 |

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 m and s categories is ±2.0, while the width of the 95% confidence intervals for the som and sbm categories is ±2.9.

Signal main effect

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | m\_ot\_x\_4 | m\_ot\_x\_5 | m\_ot\_x\_8 | m\_ot\_x\_9 | m\_po\_x\_2 | m\_po\_x\_3 |
| mean | 53.4 | 54.0 | 53.3 | 53.6 | 48.1 | 51.1 |
| N | 261 | 261 | 270 | 243 | 270 | 261 |
|  | m\_po\_x\_7 | m\_si\_x\_3 | s\_cl\_2t\_1 | s\_cl\_2t\_3 | s\_cl\_2t\_5 | s\_cl\_mt\_2 |
| mean | 51.5 | 54.2 | 51.0 | 52.5 | 59.1 | 50.6 |
| N | 270 | 252 | 270 | 261 | 270 | 270 |
|  | s\_no\_2t\_1 | s\_no\_ft\_1 | s\_no\_ft\_3 | s\_no\_mt\_1 | sbm\_js\_x\_2 | sbm\_sm\_x\_1 |
| mean | 55.0 | 46.9 | 52.6 | 50.8 | 49.8 | 56.0 |
| N | 270 | 252 | 270 | 243 | 270 | 234 |
|  | sbm\_sm\_x\_2 | sbm\_sm\_x\_6 | som\_fi\_x\_3 | som\_fi\_x\_4 | som\_ot\_x\_3 | som\_ot\_x\_5 |
| mean | 50.4 | 53.8 | 52.5 | 50.2 | 52.4 | 54.2 |
| N | 270 | 252 | 270 | 252 | 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 | Nokia | NTT-AT | T-Sys |
| mean | 63.3 | 56.7 | 47.6 | 41.3 |
| N | 1620 | 1584 | 1467 | 1611 |

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. However, the effect of this interaction compared to, say, the listener differences, the signal differences or the codec-signal interaction is relatively small. If this interaction had not been included in the statistical model, the residual standard error would have been about 8% larger.

### Post-screening of data

Of the 720 sets of 9 judgments (one for each codec, reference codec, and anchor) in this experiment, 22 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. However, in the analysis of variance, the standard error of the residuals, and thus all confidence interval widths, increases by about 1%.

# Application of Selection Rules

The Selection Rules as defined in S4-(03)0837 [6] have been applied using the data collected in the experiments being analyzed here. The following are the results.

## PSS/MMS LBRAC 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.

## PSS/MMS LBRAC Selection Rule 2

This rule ensures that each candidate codec outperforms the better of the two reference codecs in each test case. It is easiest to inspect the 8 charts above showing “all data” with confidence intervals to see which candidate codecs performed better than the reference codecs, however the confidence intervals from the ANOVA are tighter and give more statistical power. Therefore two charts are presented, one based on the Pivot Table analysis and the other on the ANOVA. Careful inspection reveals that, as expected, the differences are very minor.

The average results for each test case have been assembled in the following charts. The green cells indicate where the candidate codec is “better than” the reference codecs (in a statistical sense at the 95% level). The red cells indicate where the candidate codec is “worse than” at least one of the reference codecs (in a statistical sense at the 95% level). The light-yellow boxs indicate where the candidate codec is not statistically significantly different from the max of the two reference codecs (i.e. it is neither “better than” nor “worse than”).

**Pivot Table:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Test | Codec:  Operating condition | AAC+ | AMR-WB+ | CT | AAC | AMR-WB | Max of AAC, AMR-WB |
| A1 | 14 kbps, mono, use case A (PSS) | 50.8 | 62.6 | 51.5 | 32.9 | 44.9 | 44.9 |
| A2 | 18 kbps, stereo, use case A (PSS) | 37.5 | 55.6 | 53.3 | 20.9 | 48.2 | 48.2 |
| A3 | 24 kbps, mono, use case A (PSS) | 74.9 | 67.4 | 75.8 | 50.9 | 47.4 | 50.9 |
| A4 | 24 kbps, stereo, use case A (PSS) | 55.3 | 61.3 | 67.1 | 34.8 | 44.8 | 44.8 |
| B1 | 14 kbps, mono, use case B (MMS),  16 kHz inp. and outp. sampling rate | 45.4 | 50.7 | 44.4 | 30.7 | 46.2 | 46.2 |
| B2 | 18 kbps, stereo, use case B (MMS) | 43.3 | 50.7 | 55.7 | 22.8 | 46.8 | 46.8 |
| B3 | 14 kbps, mono, use case A (PSS),  3% FER | 43.1 | 52.5 | 44.3 | 32.1 | 24.7 | 32.1 |
| B4 | 24 kbps, stereo, use case A (PSS),  3% FER | 48.9 | 53.3 | 58.0 | 33.1 | 22.0 | 33.1 |

**ANOVA:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Test | Codec:  Operating condition | AAC+ | AMR-WB+ | CT | AAC | AMR-WB | Max of AAC, AMR-WB |
| A1 | 14 kbps, mono, use case A (PSS) | 50.8 | 62.7 | 51.6 | 32.9 | 45.0 | 45.0 |
| A2 | 18 kbps, stereo, use case A (PSS) | 37.5 | 55.6 | 53.3 | 20.9 | 48.2 | 48.2 |
| A3 | 24 kbps, mono, use case A (PSS) | 75.0 | 67.4 | 75.8 | 50.9 | 47.4 | 50.9 |
| A4 | 24 kbps, stereo, use case A (PSS) | 55.3 | 61.3 | 67.1 | 34.8 | 44.8 | 44.8 |
| B1 | 14 kbps, mono, use case B (MMS),  16 kHz inp. and outp. sampling rate | 45.5 | 50.7 | 44.4 | 30.7 | 46.2 | 46.2 |
| B2 | 18 kbps, stereo, use case B (MMS) | 43.3 | 50.7 | 55.7 | 22.8 | 46.8 | 46.8 |
| B3 | 14 kbps, mono, use case A (PSS),  3% FER | 43.1 | 52.5 | 44.4 | 32.1 | 24.7 | 32.1 |
| B4 | 24 kbps, stereo, use case A (PSS),  3% FER | 48.9 | 53.3 | 58.0 | 33.1 | 22.0 | 33.1 |

## PSS/MMS LBRAC Selection Rule 3

As described in the Selection Rules document, and clarified in document [8] the Preferred Figure of Merit calculations were performed and are presented in the following table:



# Reference Documents

1. Tdoc S4-(03)0824, “AMR-WB+ and PSS/MMS Low-Rate Audio Selection Test and Processing Plan Version 2.2.
2. RECOMMENDATION ITU-R BS.1534, Method for the subjective assessment of intermediate quality level of coding systems
3. 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.
4. Modern Applied Statistics with S, by W.N. Venables and B.D. Ripley (2002) Springer. Known colloquially as MASS.
5. MASS, p 140ff describe lm(). p 165ff describe aov(), which is a “wrapper” for lm().
6. Tdoc S4-(03)0837. PSS/MMS Audio Codec and Extended AMR-WB Selection Rules, Version 2.0.
7. Tdoc S4-(03)0433. PSS/MSS Audio Codec Selection, Design Constraints and Performance Requirements – Version 2.0.
8. Tdoc S4-040117 Implementation of the preferred FOM of PSS/MMS low-rate audio codec selection rule 3.

Annex I - Low-Rate Experiment Training and Test Items

## Training Items

The same training items were used at all test sites. They were:

m\_vo\_x\_1\_org.wav

s\_no\_ft\_9\_org.wav

sbm\_fi\_x\_9\_org.wav

som\_ot\_x\_9\_org.wav

## Test Items

|  |  |  |  |
| --- | --- | --- | --- |
| Test | Set | Item | Signal |
| A1 | a | 1 | m\_ot\_x\_8\_org.wav |
|  |  | 2 | m\_ot\_x\_a\_org.wav |
|  |  | 3 | m\_po\_x\_5\_org.wav |
|  |  | 4 | m\_po\_x\_7\_org.wav |
|  |  | 5 | s\_cl\_2t\_3\_org.wav |
|  |  | 6 | s\_cl\_2t\_4\_org.wav |
|  |  | 7 | s\_no\_2t\_1\_org.wav |
|  |  | 8 | s\_no\_ft\_1\_org.wav |
|  |  | 9 | sbm\_js\_x\_1\_org.wav |
|  |  | 10 | sbm\_ms\_x\_1\_org.wav |
|  |  | 11 | som\_fi\_x\_4\_org.wav |
|  |  | 12 | som\_ot\_x\_4\_org.wav |
|  | b | 1 | m\_ot\_x\_9\_org.wav |
|  |  | 2 | m\_ot\_x\_b\_org.wav |
|  |  | 3 | m\_po\_x\_6\_org.wav |
|  |  | 4 | m\_si\_x\_3\_org.wav |
|  |  | 5 | s\_cl\_2t\_5\_org.wav |
|  |  | 6 | s\_cl\_mt\_2\_org.wav |
|  |  | 7 | s\_no\_2t\_2\_org.wav |
|  |  | 8 | s\_no\_ft\_2\_org.wav |
|  |  | 9 | sbm\_sj\_x\_1\_org.wav |
|  |  | 10 | sbm\_sm\_x\_6\_org.wav |
|  |  | 11 | som\_ot\_x\_5\_org.wav |
|  |  | 12 | som\_ot\_x\_6\_org.wav |
|  |  |  |  |
| Test | Set | Item | Signal |
| A2 | a | 1 | m\_ot\_x\_4\_org.wav |
|  |  | 2 | m\_ot\_x\_5\_org.wav |
|  |  | 3 | m\_po\_x\_2\_org.wav |
|  |  | 4 | m\_po\_x\_3\_org.wav |
|  |  | 5 | s\_cl\_2t\_4\_org.wav |
|  |  | 6 | s\_cl\_ft\_3\_org.wav |
|  |  | 7 | s\_no\_2t\_3\_org.wav |
|  |  | 8 | s\_no\_mt\_1\_org.wav |
|  |  | 9 | sbm\_js\_x\_1\_org.wav |
|  |  | 10 | sbm\_sm\_x\_4\_org.wav |
|  |  | 11 | som\_fi\_x\_3\_org.wav |
|  |  | 12 | som\_ot\_x\_2\_org.wav |
|  | b | 1 | m\_ot\_x\_6\_org.wav |
|  |  | 2 | m\_ot\_x\_7\_org.wav |
|  |  | 3 | m\_po\_x\_4\_org.wav |
|  |  | 4 | m\_si\_x\_2\_org.wav |
|  |  | 5 | s\_cl\_2t\_5\_org.wav |
|  |  | 6 | s\_cl\_mt\_2\_org.wav |
|  |  | 7 | s\_no\_ft\_3\_org.wav |
|  |  | 8 | s\_no\_ft\_4\_org.wav |
|  |  | 9 | sbm\_js\_x\_2\_org.wav |
|  |  | 10 | sbm\_sm\_x\_5\_org.wav |
|  |  | 11 | som\_ad\_x\_1\_org.wav |
|  |  | 12 | som\_ot\_x\_3\_org.wav |
|  |  |  |  |
| Test | Set | Item | Signal |
| A3 | a | 1 | m\_ot\_x\_2\_org.wav |
|  |  | 2 | m\_po\_x\_1\_org.wav |
|  |  | 3 | m\_po\_x\_2\_org.wav |
|  |  | 4 | m\_si\_x\_1\_org.wav |
|  |  | 5 | s\_cl\_2t\_1\_org.wav |
|  |  | 6 | s\_cl\_ft\_3\_org.wav |
|  |  | 7 | s\_no\_2t\_1\_org.wav |
|  |  | 8 | s\_no\_mt\_1\_org.wav |
|  |  | 9 | sbm\_ms\_x\_1\_org.wav |
|  |  | 10 | sbm\_sm\_x\_2\_org.wav |
|  |  | 11 | som\_nt\_x\_1\_org.wav |
|  |  | 12 | som\_ot\_x\_2\_org.wav |
|  | b | 1 | m\_ot\_x\_3\_org.wav |
|  |  | 2 | m\_po\_x\_3\_org.wav |
|  |  | 3 | m\_po\_x\_4\_org.wav |
|  |  | 4 | m\_si\_x\_2\_org.wav |
|  |  | 5 | s\_cl\_2t\_2\_org.wav |
|  |  | 6 | s\_cl\_mt\_2\_org.wav |
|  |  | 7 | s\_no\_2t\_2\_org.wav |
|  |  | 8 | s\_no\_ft\_4\_org.wav |
|  |  | 9 | sbm\_js\_x\_2\_org.wav |
|  |  | 10 | sbm\_sm\_x\_5\_org.wav |
|  |  | 11 | som\_ot\_x\_1\_org.wav |
|  |  | 12 | som\_ot\_x\_3\_org.wav |
|  |  |  |  |
| Test | Set | Item | Signal |
| A4 | a | 1 | m\_ch\_x\_1\_org.wav |
|  |  | 2 | m\_ot\_x\_2\_org.wav |
|  |  | 3 | m\_po\_x\_1\_org.wav |
|  |  | 4 | m\_si\_x\_1\_org.wav |
|  |  | 5 | s\_cl\_2t\_1\_org.wav |
|  |  | 6 | s\_cl\_mt\_1\_org.wav |
|  |  | 7 | s\_no\_2t\_1\_org.wav |
|  |  | 8 | s\_no\_ft\_2\_org.wav |
|  |  | 9 | sbm\_ms\_x\_1\_org.wav |
|  |  | 10 | sbm\_sm\_x\_2\_org.wav |
|  |  | 11 | som\_fi\_x\_1\_org.wav |
|  |  | 12 | som\_nt\_x\_1\_org.wav |
|  | b | 1 | m\_cl\_x\_1\_org.wav |
|  |  | 2 | m\_cl\_x\_2\_org.wav |
|  |  | 3 | m\_ot\_x\_1\_org.wav |
|  |  | 4 | m\_ot\_x\_3\_org.wav |
|  |  | 5 | s\_cl\_2t\_2\_org.wav |
|  |  | 6 | s\_cl\_2t\_3\_org.wav |
|  |  | 7 | s\_no\_2t\_2\_org.wav |
|  |  | 8 | s\_no\_ft\_1\_org.wav |
|  |  | 9 | sbm\_sm\_x\_1\_org.wav |
|  |  | 10 | sbm\_sm\_x\_3\_org.wav |
|  |  | 11 | som\_fi\_x\_2\_org.wav |
|  |  | 12 | som\_ot\_x\_1\_org.wav |
|  |  |  |  |
| Test | Set | Item | Signal |
| B1 | a | 1 | m\_ch\_x\_1\_org.wav |
|  |  | 2 | m\_cl\_x\_1\_org.wav |
|  |  | 3 | m\_po\_x\_2\_org.wav |
|  |  | 4 | m\_po\_x\_3\_org.wav |
|  |  | 5 | s\_cl\_2t\_4\_org.wav |
|  |  | 6 | s\_cl\_ft\_3\_org.wav |
|  |  | 7 | s\_no\_2t\_2\_org.wav |
|  |  | 8 | s\_no\_ft\_2\_org.wav |
|  |  | 9 | sbm\_js\_x\_2\_org.wav |
|  |  | 10 | sbm\_ms\_x\_1\_org.wav |
|  |  | 11 | som\_ad\_x\_1\_org.wav |
|  |  | 12 | som\_ot\_x\_4\_org.wav |
|  | b | 1 | m\_po\_x\_4\_org.wav |
|  |  | 2 | m\_po\_x\_5\_org.wav |
|  |  | 3 | m\_po\_x\_6\_org.wav |
|  |  | 4 | m\_si\_x\_2\_org.wav |
|  |  | 5 | s\_cl\_2t\_2\_org.wav |
|  |  | 6 | s\_cl\_mt\_1\_org.wav |
|  |  | 7 | s\_no\_2t\_3\_org.wav |
|  |  | 8 | s\_no\_ft\_4\_org.wav |
|  |  | 9 | sbm\_sj\_x\_1\_org.wav |
|  |  | 10 | sbm\_sm\_x\_4\_org.wav |
|  |  | 11 | som\_ot\_x\_2\_org.wav |
|  |  | 12 | som\_ot\_x\_6\_org.wav |
|  |  |  |  |
| Test | Set | Item | Signal |
| B2 | a | 1 | m\_cl\_x\_2\_org.wav |
|  |  | 2 | m\_ot\_x\_1\_org.wav |
|  |  | 3 | m\_ot\_x\_8\_org.wav |
|  |  | 4 | m\_ot\_x\_a\_org.wav |
|  |  | 5 | s\_cl\_2t\_4\_org.wav |
|  |  | 6 | s\_cl\_2t\_5\_org.wav |
|  |  | 7 | s\_no\_2t\_3\_org.wav |
|  |  | 8 | s\_no\_ft\_2\_org.wav |
|  |  | 9 | sbm\_js\_x\_1\_org.wav |
|  |  | 10 | sbm\_sm\_x\_4\_org.wav |
|  |  | 11 | som\_fi\_x\_1\_org.wav |
|  |  | 12 | som\_ot\_x\_5\_org.wav |
|  | b | 1 | m\_ch\_x\_1\_org.wav |
|  |  | 2 | m\_cl\_x\_1\_org.wav |
|  |  | 3 | m\_ot\_x\_9\_org.wav |
|  |  | 4 | m\_ot\_x\_b\_org.wav |
|  |  | 5 | s\_cl\_2t\_3\_org.wav |
|  |  | 6 | s\_cl\_mt\_1\_org.wav |
|  |  | 7 | s\_no\_ft\_1\_org.wav |
|  |  | 8 | s\_no\_ft\_3\_org.wav |
|  |  | 9 | sbm\_sm\_x\_1\_org.wav |
|  |  | 10 | sbm\_sm\_x\_3\_org.wav |
|  |  | 11 | som\_fi\_x\_2\_org.wav |
|  |  | 12 | som\_ot\_x\_6\_org.wav |
|  |  |  |  |
| Test | Set | Item | Signal |
| B3 | a | 1 | m\_cl\_x\_1\_org.wav |
|  |  | 2 | m\_ot\_x\_1\_org.wav |
|  |  | 3 | m\_ot\_x\_5\_org.wav |
|  |  | 4 | m\_ot\_x\_7\_org.wav |
|  |  | 5 | s\_cl\_2t\_1\_org.wav |
|  |  | 6 | s\_cl\_2t\_3\_org.wav |
|  |  | 7 | s\_no\_2t\_3\_org.wav |
|  |  | 8 | s\_no\_ft\_4\_org.wav |
|  |  | 9 | sbm\_js\_x\_2\_org.wav |
|  |  | 10 | sbm\_ms\_x\_1\_org.wav |
|  |  | 11 | som\_ad\_x\_1\_org.wav |
|  |  | 12 | som\_fi\_x\_2\_org.wav |
|  | b | 1 | m\_ch\_x\_1\_org.wav |
|  |  | 2 | m\_cl\_x\_2\_org.wav |
|  |  | 3 | m\_ot\_x\_4\_org.wav |
|  |  | 4 | m\_ot\_x\_6\_org.wav |
|  |  | 5 | s\_cl\_2t\_2\_org.wav |
|  |  | 6 | s\_cl\_mt\_1\_org.wav |
|  |  | 7 | s\_no\_ft\_3\_org.wav |
|  |  | 8 | s\_no\_mt\_1\_org.wav |
|  |  | 9 | sbm\_js\_x\_1\_org.wav |
|  |  | 10 | sbm\_sm\_x\_1\_org.wav |
|  |  | 11 | som\_fi\_x\_1\_org.wav |
|  |  | 12 | som\_fi\_x\_3\_org.wav |
|  |  |  |  |
| Test | Set | Item | Signal |
| B4 | a | 1 | m\_ot\_x\_4\_org.wav |
|  |  | 2 | m\_ot\_x\_8\_org.wav |
|  |  | 3 | m\_po\_x\_2\_org.wav |
|  |  | 4 | m\_po\_x\_7\_org.wav |
|  |  | 5 | s\_cl\_2t\_1\_org.wav |
|  |  | 6 | s\_cl\_2t\_5\_org.wav |
|  |  | 7 | s\_no\_2t\_1\_org.wav |
|  |  | 8 | s\_no\_ft\_3\_org.wav |
|  |  | 9 | sbm\_js\_x\_2\_org.wav |
|  |  | 10 | sbm\_sm\_x\_2\_org.wav |
|  |  | 11 | som\_fi\_x\_3\_org.wav |
|  |  | 12 | som\_ot\_x\_5\_org.wav |
|  | b | 1 | m\_ot\_x\_5\_org.wav |
|  |  | 2 | m\_ot\_x\_9\_org.wav |
|  |  | 3 | m\_po\_x\_3\_org.wav |
|  |  | 4 | m\_si\_x\_3\_org.wav |
|  |  | 5 | s\_cl\_2t\_3\_org.wav |
|  |  | 6 | s\_cl\_mt\_2\_org.wav |
|  |  | 7 | s\_no\_ft\_1\_org.wav |
|  |  | 8 | s\_no\_mt\_1\_org.wav |
|  |  | 9 | sbm\_sm\_x\_1\_org.wav |
|  |  | 10 | sbm\_sm\_x\_6\_org.wav |
|  |  | 11 | som\_fi\_x\_4\_org.wav |
|  |  | 12 | som\_ot\_x\_3\_org.wav |