INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11
MPEG2011/N12232
July 2011, Torino, Italy

Title: **USAC Verification Test Report**
Source: **Audio subgroup**

Executive summary

The Unified Speech and Audio Coding (USAC) work item progressed to FDIS stage in July, 2011 and is expected to issue as ISO/IEC 23003-3, Unified Speech and Audio Coding in 2011. Verification tests were conducted to assess the subjective quality of this new specification. The tests included speech, music and speech mixed with music, and covered a large range of bit rates from 8 to 96 kbps as well as both mono and stereo operating points. Depending on the listening test, 6 to 13 test sites participated with a commitment to use at least 8 listeners per test. This resulted in a large data set of more than 38000 individual scores.

The statistical analysis of the test data resulted in the following conclusions:

- The goal of the work item was achieved, that is to produce a single coding technology which is always at least as good as the better of the state-of-the-art coders, namely HE-AACv2 and AMR-WB+, on any signal content type (speech, music and speech mixed with music).
- In fact, when considering any of the three content types, and of course for all signal content types grouped together, USAC is always better than the best codec tailored for general audio or speech (i.e. HE-AACv2 and AMR-WB+). This holds true for the whole range of bit rates tested from 8 kbps mono to 96 kbps stereo. For most bit rates tested this performance advantage is large.
- USAC provides a much more consistent quality across all signal content types than the other systems tested.

1 Introduction

The Unified Speech and Audio Coding (USAC) technology was developed to provide coding of signals having an arbitrary mix of speech and audio content with consistent quality across content types, in particular at intermediate and low bit rates. From the beginning of the USAC work item, it has been envisioned that a new audio codec providing a substantial quality gain over existing audio compression technology would be of particular value on operations at medium bit rates, for example between 24 and 32 kbps. Potential applications include Digital Radio, Mobile TV, Audio books, Multimedia Download and Real-time Play on Mobile devices. While the goal is to ensure suitable quality at medium bit rates, it is also important to preserve quality over the lower bit rate range (e.g. down to 12 kbps and even below) for applications with very demanding bit rate limitations, and that quality scales up at higher rates (up to 64 kbps and beyond).

The objectives of the USAC work item were very ambitious. It was recognized that no single technology could provide consistently good quality for all signal content types (speech, music, speech mixed with music) especially at low bit rates (typically below 32 kbps). Hence, a so-called “Virtual Codec” (VC) was defined as the technology with the target quality to match or exceed. This VC, which actually does not exist as a single technology, was defined as the best of HE-AACv2 or AMR-WB+ on a given test item at a given operating point (bit rate and number of audio channels). The VC has the advantage that the scores on a given item are selected a posteriori from the best performance of two independent state-of-the-art codecs (HE-AACv2 and AMR-WB+). The desired New Technology (NT) from the USAC work item was required

to match or exceed the quality of the VC all on its own as a single technology. At the conclusion of the standardization process, NT became USAC.

The Requirements for the work on Unified Speech and Audio Coding were as follows:

- That the performance of NT shall not be worse than the performance of VC when both are operated at the same bitrate.
- Although the performance of NT is most important at low bitrates, e.g. below 24kb/s/channel, it shall not be worse than the performance of current MPEG technology when both are operated at higher bitrates.

USAC progressed to FDIS stage in July, 2011 and is expected to issue as ISO/IEC 23003-3, Unified Speech and Audio Coding in 2011. In order to permit National Bodies and industry to evaluate the performance of the USAC specification, verification tests have been conducted in June 2011. The verification tests were designed to provide information on the subjective performance of USAC in mono and stereo over a wide range of bit rates from 8 to 96 kbps. This document reports the results of the verification tests.

2 Listening tests description

Three listening tests have been performed. The conditions included in each test are given in the tables below. Along with USAC, two other audio coding standards were included as references, namely HE-AACv2 and AMR-WB+. These two reference codecs form what was termed the Virtual Codec (VC), which is defined as the best of either HE-AACv2 or AMR-WB+ on a given test item at a given operating point (bit rate and number of audio channels). The aim of the Verification Tests was to verify that USAC is as good as or better than the VC over all tested operating points.

Conditions for Test 1 (mono at low rates)

Condition	**Label**
Hidden reference	HR
Low pass anchor at 3.5 kHz	LP3500
Low pass anchor at 7 kHz	LP7000
AMR-WB+ at 8 kbps	AMR-8
AMR-WB+ at 12 kbps	AMR-12
AMR-WB+ at 24 kbps	AMR-24
HE-AAC v2 at 12 kbps	HE-AAC-12
HE-AAC v2 at 24 kbps	HE-AAC-24
USAC at 8 kbps	USAC-8
USAC at 12 kbps	USAC-12
USAC at 16 kbps	USAC-16
USAC at 24 kbps	USAC-24

Conditions for Test 2 (stereo at low rates)

Condition	**Label**
Hidden reference	HR
Low pass anchor at 3.5 kHz*	LP3500
Low pass anchor at 7 kHz*	LP7000
AMR-WB+ at 16 kbps	AMR-16
AMR-WB+ at 24 kbps	AMR-24
HE-AAC v2 at 16 kbps	HE-AAC-16
HE-AAC v2 at 24 kbps	HE-AAC-24
USAC at 16 kbps	USAC-16
USAC at 20 kbps	USAC-20
USAC at 24 kbps	USAC-24

Conditions for Test 3 (stereo at high rates)

Condition	Label
Hidden reference	HR
Low pass anchor at 3.5 kHz*	LP3500
Low pass anchor at 7 kHz*	LP7000
AMR-WB+ at 32 kbps	AMR-32
HE-AAC v2 at 32 kbps	HE-AAC-32
HE-AAC v2 at 64 kbps	HE-AAC-64
HE-AAC v2 at 96 kbps	HE-AAC-96
USAC at 32 kbps	USAC-32
USAC at 48 kbps	USAC-48
USAC at 64 kbps	USAC-64
USAC at 96 kbps	USAC-96

*Bandlimited but keeping the same stereo width as the original (hidden reference).

3 Test items

The table below gives the list of test and training items used in the verification tests of USAC. The 24 test items and 3 training items have been selected by an expert panel formed by David Virette (Huawei Technologies Co., Ltd.), Roch Lefebvre (VoiceAge Corporation), Toru Chinen (Sony Corporation) and Max Neuendorf (Fraunhofer IIS). The columns in the table give respectively the item name, item category, a short description of the item and the allocation of each test item to a subtest such that half of the listeners in a test could listen to the items in subtest 1, and the other half of the listeners could listen to the items in subtest 2.

The expert panel made a selection of test items using the following guidelines:

- Make sure the test items, in each category, cover content representative of the envisioned application scenarios (for example not just items like castanets and pitch pipe which cover very specific audio content behaviour)
- Select unbiased items, i.e. items that are critical for each of the systems in the test (i.e. USAC and all the reference codecs used)
- Take into account the following application scenarios for USAC
    - Digital Radio, Mobile TV, Audio books focusing on speech and speech with background noise contents including announcement, advertisement, and narration.
    - Multimedia Download and Real-time Play on Mobile devices focusing on various types of Music and movie contents

Item Name	Cat.	Description	Subtest
es03	speech	English female	1
es02	speech	German male	1
Arirang_speech	speech	Korean male	1
PannedSpeechEkonomiekot_cut	speech	Swedish male	1
es01_short	speech	English female	2
te1_mg54_speech	speech	German male	2
Green_speech	speech	Korean male	2
te19_short	speech	English male	2
carrot_speech_short	speech	English male	Training
Alice_short	mixed	English female between/over classical music	1
lion	mixed	English male between effects	1

noodleking_short	mixed	English male between music (advert)	1
SpeechOverMusic_1_short	mixed	English female over noise	1
dora_short	mixed	English female between/over effects	2
mod101	mixed	French speaker over noise	2
SpeechOverMusic_2_short	mixed	English female speech over music	2
SpeechOverMusic_3_short	mixed	English male speech over music/noise (advert)	2
SpeechOverMusic_5_short	mixed	English male speech over music (advert)	Training
Schmoo_lvl2	music	Electronic Pop Music	1
trilogy_short	music	Saxophone melody with accompaniment	1
te09_short	music	English female over music	1
id4	music	Emphatic movie score	1
Siefried02	music	Classical music, melodic	2
taurusb	music	Male singing over rock music	2
phi3_short	music	Pop music (instrumental)	2
phi4_short	music	English male over music	2
dongwoo_short	music	Korean female over music	Training

4 Test plan

Test methodology

All tests used the MUSHRA method described in [1]. A quality scale is used where the intervals are labeled "bad", "poor", "fair", "good" and "excellent". The subjective response is recorded on a scale ranging from 0 to 100, with no decimals digits.

In the listening test methodology, the following were incorporated:

- Test items of about 8 seconds in duration
- At least 20 listeners in total per test (counting all the listeners from all test sites participating at a given test)
- It is strongly recommended that there be at least 8 listeners per test per site
- Expert listeners are preferred
- In a given test at a given site, the same set of listeners shall listen to and rate all systems under test for all items presented to the listeners.
    - Note that in order to increase the number of test items in each of the three content categories (speech, music and speech mixed with music), half the listeners in a test will listen to half of the test items, and the other half of the listeners will listen to the other half of the test items. Specifically, since there are 8 test items per category, half of the listeners will listen to the first 4 test items in each content category and for each system under test, and the other half of the listeners will listen to the last 4 test items in each content category and for each system under test.
- Listeners shall receive training before participating in each test
- 3 training items, one per content category, have also been selected by the panel of audio experts. These shall be used as “training tests” to be done before each of the tests using the other 8 test items per content category.
- Computer-based MUSHRA presentation shall be used which supports instantaneous switching between the test items
- Tests shall be divided into sessions of e.g. 20 minutes length in order to prevent listener fatigue
- Only one listener at a time due to open headphones and interactive control
- Tests shall be conducted in acoustically isolated environments, such as a commercial sound booth.
- Reference quality headphones shall be used (e.g. Sennheiser HD600 or STAX for headphone listening test.
- Test sites are requested to report the listening setup actually used

Post-screening of listener responses should be applied as follows. If, for any test item in a given test, either of the following criterion are not satisfied:

- The listener score for the hidden reference is greater than or equal to 90 (i.e. HR >= 90)
- The listener scores the hidden reference, the 7.0 kHz lowpass anchor and the 3.5 kHz lowpass anchor are monotonically decreasing (i.e. HR >= LP70 >= LP35).

Then all listener responses in that test are removed from consideration.

Test item preparation

4.1.1 Preparation of original items

The Test Administrator (Schuyler Quackenbush) prepared the original items. All original items listed in the table in Section 0

- Have been adjusted in level as indicated in the "level" column in the table listing test items
- Have been limited in duration to approximately 8 seconds
- Have been trimmed to have less than 10 ms of initial and trailing silence
- Have gradual fade-in and gradual fade out
- Have monophonic versions derived from stereophonic versions as per the information provided by the party having sent the original item to the Test Administrator

All items were concatenated to form a single file. One second of silence was inserted after the training items.

4.1.2 Preparation of test items

The Test Administrator prepared the original (uncoded) test items before submitting them to FhG/VoiceAge/Dolby for encoding by the different systems under test (USAC, AMR-WB+ and HE-AACv2). The USAC encoded/decoded files were produced using the Baseline USAC encoder/decoder which excludes the following tools: time-warped filterbank, DFT-based harmonic transposer and fractional delay decorrelator.

Test sites

The following table indicates (with an "X") the test sites that participated in each of the tests. As strongly advised in the test plan, each test sites provided data from at least 8 listeners in each test it contributed to.

Test	VoiceAge Corporation	Fraunhofer IIS	Dolby	Huawei	Orange Labs	Sony Corporation	Samsung	Philips	Panasonic
1	X	X	X	X		X	X		X
2	X	X	X		X	X	X	X	X
3	X	X	X		X	X	X		

Test	ETRI	Ericsson	DOCOMO	Yonsei University	Telcordia and PUT	Qualcomm
1	X	X	X	X	X	X
2						
3						

5 Test Results

Listener post-screening

The combined number of listeners from all test sites in each test was as follows:

Test 1, subtest 1: 68

Test 1, subtest 2: 70
Test 2, subtest 1: 47
Test 2, subtest 2: 45
Test 3, subtest 1: 29
Test 3, subtest 2: 29

After applying listener post-screening according to section 5.1, the number of remaining listeners whose raw data were used for the statistical analysis was as follows:

Test 1, subtest 1: 62
Test 1, subtest 2: 62
Test 2, subtest 1: 41
Test 2, subtest 2: 37
Test 3, subtest 1: 26
Test 3, subtest 2: 27

Statistical analysis

This section presents the statistical analysis of the test results. Figures 1 to 3 show the average absolute scores for each codec (including the VC – Virtual Codec defined in section 3) at the different operating points tested. The scores from all test sites, after listener post-screening, are pooled for this analysis. Vertical bars around each average score indicate the 95% confidence intervals using a t-distribution. The vertical axis in Figures 1 to 3 uses the MUSHRA voting scale:

Value between 0 and 20:	BAD
Value between 21 and 40:	POOR
Value between 41 and 60:	FAIR
Value between 61 and 80:	GOOD
Value between 81 and 100:	EXCELLENT

Note that the mean values and confidence intervals used for the statistical analysis in this section are included in file “MeanValues.xls” which is included in the ZIP file from which this contribution was retrieved.

It should also be noted that the results of the three tests must be analyzed independently and must not be compared against each other, because the raw scores from which the results have been computed were produced from independently conducted tests.

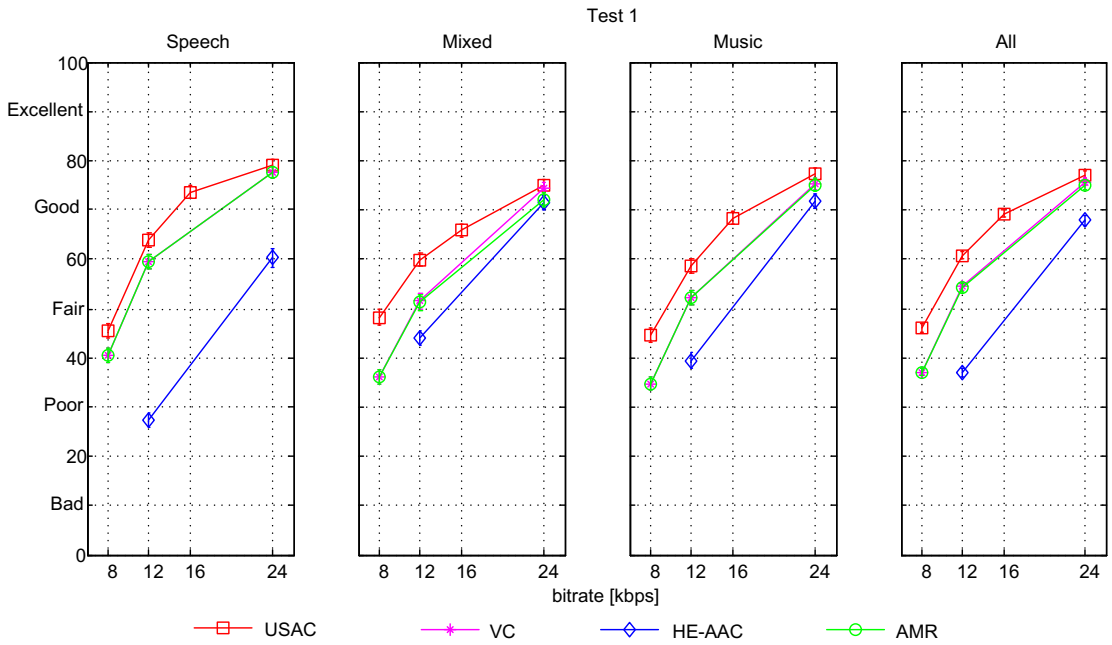


Figure 1. Average absolute scores in Test 1 for USAC, HE-AACv2 (HE-AAC in the legend), AMR-WB+ (AMR in the legend) and the Virtual Codec (VC)

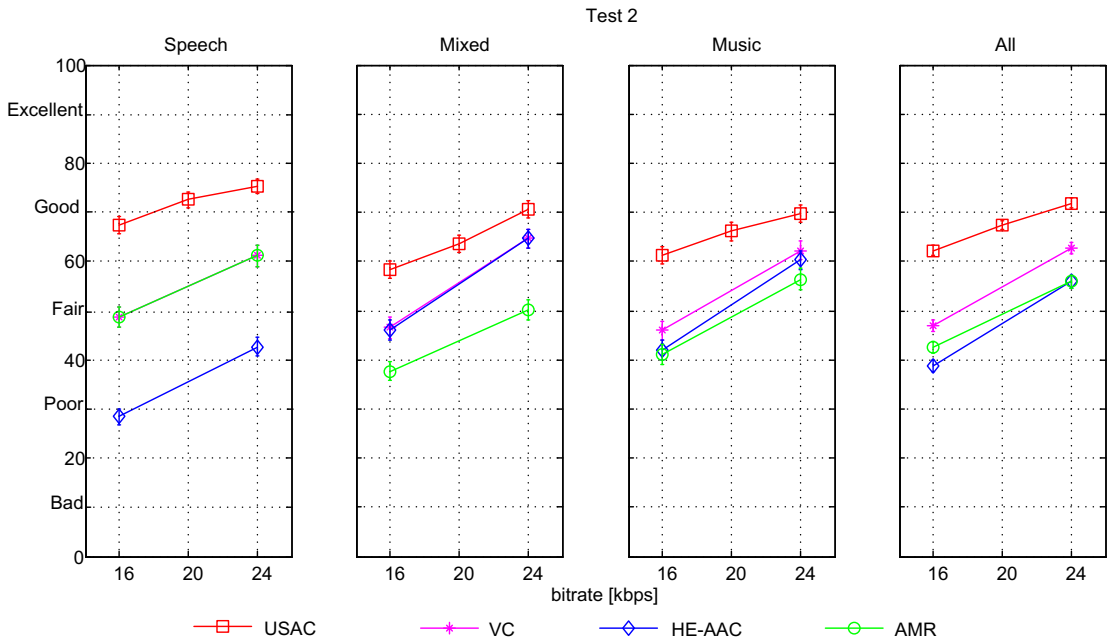


Figure 2. Average absolute scores in Test 2 for USAC, HE-AACv2 (HE-AAC in the legend), AMR-WB+ (AMR in the legend) and the Virtual Codec (VC)

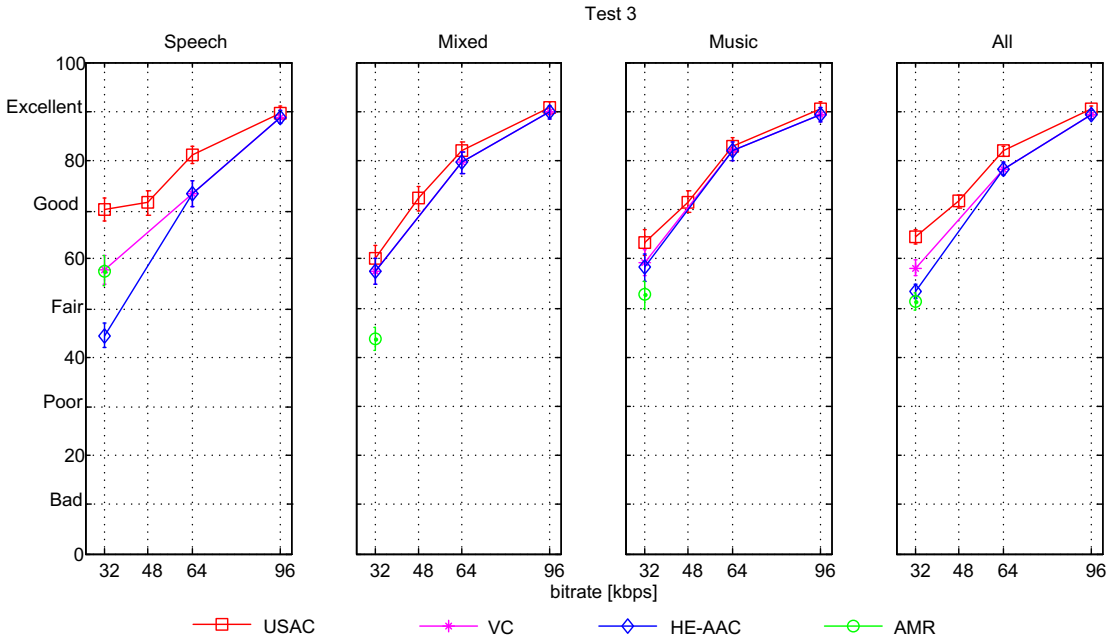


Figure 3. Average absolute scores in Test 3 for USAC, HE-AACv2 (HE-AAC in the legend), AMR-WB+ (AMR in the legend) and the Virtual Codec (VC)

Figures 1 to 3 show that, when averaging over all content types, the average absolute score of USAC is significantly above that of the VC, with 95% confidence intervals not overlapping by a large margin. Two exceptions are at 24 kbps mono and 96 kbps stereo where USAC and the VC have overlapping confidence intervals, but with the average score of USAC numerically above that of the VC. Furthermore, Figures 1 to 3 show that when considering each signal content type individually (speech, music or speech mixed with music), the absolute score for USAC is always greater than the absolute score of the VC, and often by a large margin. This is most apparent in Test 2 (stereo operation between 16 and 24 kbps), with a 5 to 20 point advantage for USAC on the 100-point scale. A third observation from Figures 1 to 3 is that the quality for USAC is much more consistent across signal content types than the two state-of-the-art codecs considered (HE-AACv2 and AMR-WB+). This is especially apparent at medium and low rate operation (Figures 1 and 2).

Figures 4 to 6 show the differential scores analysis for each test. The scores shown are the averages of the score differences given by each listener to USAC and the VC. The vertical lines show the 95% confidence intervals. The average differential scores are shown, for each bit rate tested, per signal content type and overall all test items. Recall there were 8 speech items, 8 music items and 8 speech-mixed-with- music items in the tests, hence a total of 24 test items. If the average differential score is above 0 with the 95% confidence interval not crossing 0, this implies that USAC is significantly better than the VC in the statistical sense. Figures 4 to 6 show that the differential score is never below 0, and that in most cases the differential score is greater than the 95% confidence interval by a large margin. From this, it can be concluded that USAC not only meets the requirement as outlined in the Call for Proposals but actually exceeds it in many cases by providing significantly better quality than the better of HE-AACv2 or AMR-WB+.

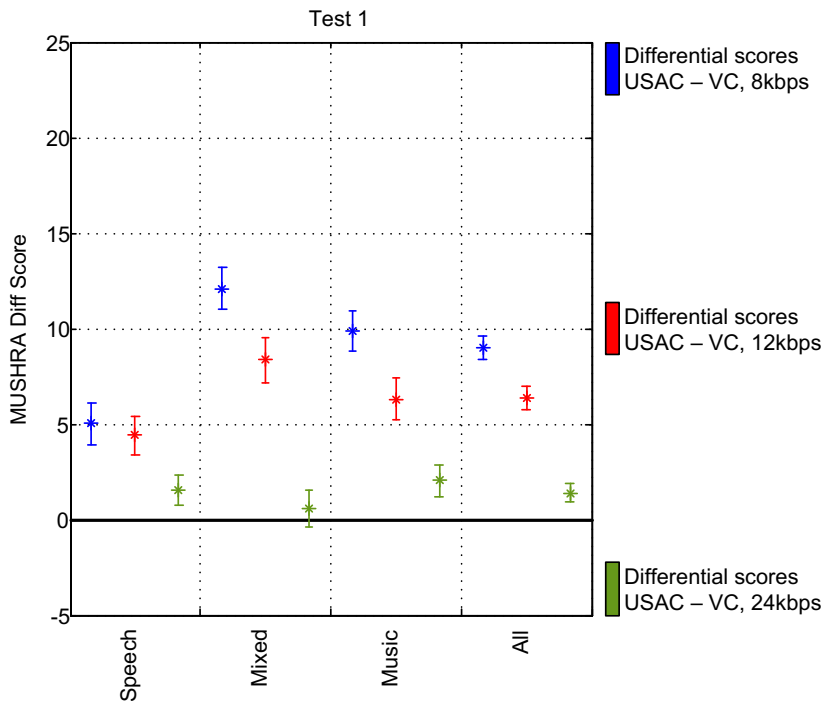


Figure 4. Average differential scores (USAC – VC) in Test 1 per content type and over all items, with 95% confidence intervals.

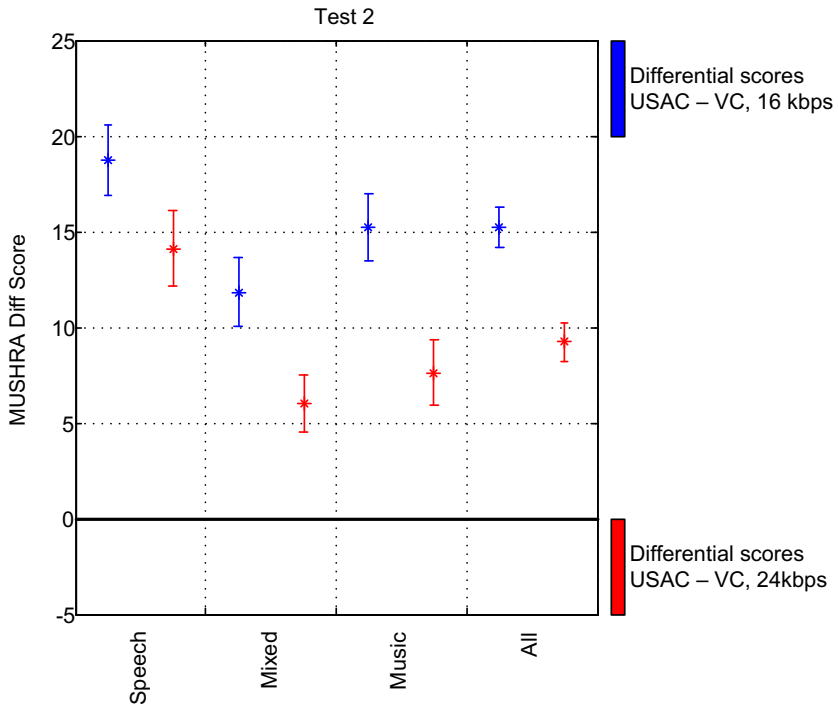


Figure 5. Average differential scores (USAC – VC) in Test 2 per content type and over all items, with 95% confidence intervals.

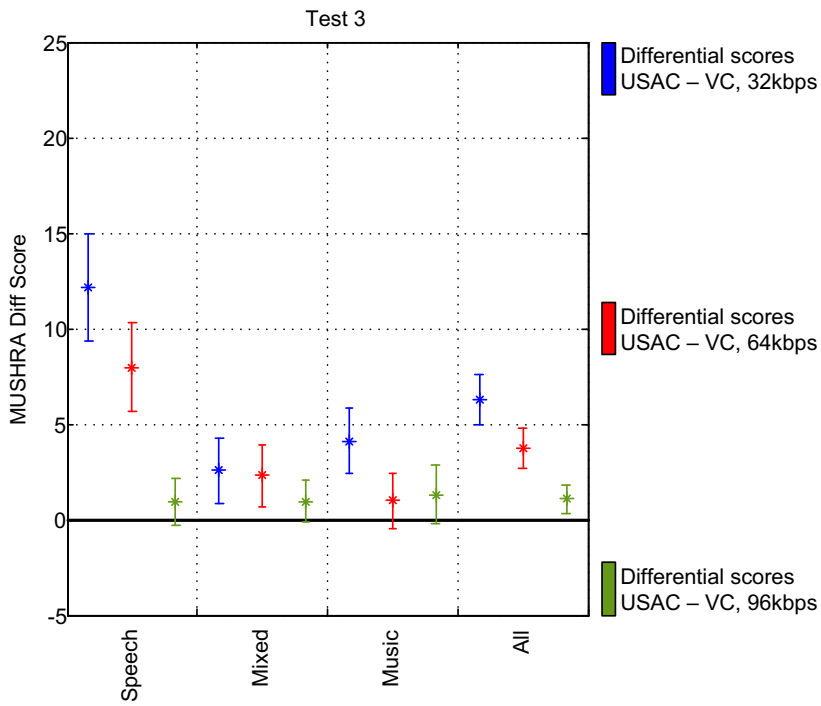


Figure 6. Average differential scores (USAC – VC) in Test 3 per content type and over all items, with 95% confidence intervals.

As applied to Figures 4 to 6, the minimum requirement of the USAC work item was that no differential score (USAC –VC) should be below 0 by more than the 95% confidence interval. Figures 4 to 6 show that not only this requirement is met but that actually it is significantly exceeded in most tested cases by pushing further the sound quality at low and mid rates (Figures 1 and 2) and even at high rates up to 64 kbps stereo for speech and mixed content.

6 Conclusion

The USAC work item aimed at producing a single technology capable of handling speech, music and speech mixed with music signals with quality always at least as good as the best of two state-of-the-art reference codecs, one optimized for speech and mixed content (AMR-WB+) and the other optimized for music and general audio (HE-AACv2). To meet the needs of several applications such as Digital Radio, Mobile TV, Audio books, Multimedia Download and Real-time Play on Mobile devices, bit rates of 32 kbps and less have been particularly targeted since it had been recognized that no single technology (and in particular neither AMR-WB+ nor HE-AACv2) could meet these quality objectives. It was also required that the quality of this new technology scales up at higher rates (up to 64 kbit/s and beyond).

The statistical analysis of the USAC verification test data presented in this report shows that these objectives have not only been met but actually exceeded. Not only does USAC at least matches the quality of the best of HE-AACv2 and AMR-WB+ on all signal content types and at all bit rates tested (from 8 mono to 96 kbps stereo,), but USAC actually pushes further the sound quality, and often by a large margin, in the bit rate range from 8 kbps mono to 64 kbps stereo.

7 References

[1] ITU-R Recommendation BS.1534-1, “Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)”, International Telecommunications Union, Geneva, Switzerland, 2001

Annex 1 Statistical Analysis Details

1.1 Means and Confidence Intervals per Category and Overall

This section presents tables of mean score and 95% confidence intervals for each test. For each table, the test is indicated in the upper left corner. The leftmost column indicates system under test. The next four columns give the mean sores: the first three for categories speech, music, mixed and the fourth with all categories combined. The final four columns give the "± value" of the 95% Confidence Interval (CI): the first three for categories speech, music, mixed and the fourth with all categories combined. Hence, the 95% Confidence Interval on the mean is constructed as

95% Confidence Interval = mean ± CI value

The last rows in each table pertain to the Virtual Codec (VC). The first set give the VC mean score at each tested bit rate. The last set gives the difference between the mean score of USAC and VC at each tested bit rate.

Test1	Mean Scores				Confidence Interval Values			
	speech	music	mixed	all	speech	music	mixed	all
USAC-08	45.394	48.122	44.518	46.012	1.491	1.578	1.483	0.878
USAC-12	63.882	59.864	58.547	60.764	1.413	1.405	1.421	0.822
USAC-16	73.622	65.809	68.372	69.268	1.335	1.360	1.410	0.806
USAC-24	79.077	74.850	77.242	77.056	1.190	1.328	1.228	0.726
HE-AAC-12	27.159	43.945	39.358	36.820	1.399	1.476	1.561	0.926
HE-AAC-24	60.161	71.467	71.685	67.771	1.786	1.391	1.394	0.926
AMR-08	40.404	36.035	34.675	37.038	1.432	1.439	1.326	0.816
AMR-12	59.449	51.140	52.240	54.276	1.538	1.567	1.508	0.906
AMR-24	77.535	71.984	75.057	74.858	1.202	1.401	1.281	0.756
3.5LP	22.620	21.463	21.439	21.841	0.904	0.901	0.837	0.509
7.0LP	46.581	42.644	43.187	44.138	1.502	1.398	1.384	0.828
original	99.919	99.896	99.931	99.915	0.063	0.076	0.063	0.039
VC-08	40.404	36.035	34.675	37.038	1.432	1.439	1.326	0.816
VC-12	59.449	51.502	52.240	54.397	1.538	1.557	1.508	0.903
VC-24	77.535	74.276	75.199	75.670	1.202	1.341	1.254	0.733
USAC-VC-08	4.990	12.087	9.843	8.974	1.097	1.131	1.053	0.648
USAC-VC-12	4.433	8.362	6.307	6.367	1.026	1.166	1.102	0.639
USAC-VC-24	1.543	0.573	2.043	1.386	0.791	0.956	0.864	0.504

Test2	Mean Scores				Confidence Interval Values			
	speech	music	mixed	all	speech	music	mixed	all
USAC-16	67.484	58.314	61.010	62.269	1.691	1.838	1.733	1.039
USAC-20	72.538	63.529	66.202	67.423	1.519	1.767	1.786	1.005
USAC-24	75.394	70.615	69.734	71.915	1.543	1.716	1.754	0.976
AMR-16	48.631	37.606	40.913	42.384	1.663	2.024	1.959	1.192
AMR-24	61.417	50.054	56.186	55.886	1.930	1.901	1.976	1.272
HE-AAC-16	28.218	46.151	41.830	38.733	2.066	1.892	1.946	1.171
HE-AAC-24	42.519	64.628	60.397	55.848	2.167	2.038	2.165	1.258
3.5LP	21.205	20.147	19.487	20.280	1.222	1.214	1.102	0.681
7.0LP	42.032	39.058	39.173	40.088	1.906	1.707	1.813	1.045
Orig	99.907	99.910	99.946	99.921	0.084	0.084	0.061	0.044
VC-16	48.631	46.497	45.782	46.970	2.066	2.087	1.876	1.160
VC-24	61.417	64.628	62.240	62.762	2.167	1.901	1.911	1.152
USAC-VC-16	18.853	11.817	15.228	15.299	1.884	1.802	1.723	1.054
USAC-VC-24	13.978	5.987	7.494	9.153	1.937	1.506	1.688	1.014

Test3	Mean Scores				Confidence Interval Values			
	speech	music	mixed	all	speech	music	mixed	all
USAC-32	70.019	59.953	63.373	64.448	2.364	2.682	2.578	1.497
USAC-48	71.387	72.278	71.566	71.744	2.402	2.401	2.246	1.350
USAC-64	81.217	81.877	82.858	81.984	1.803	1.907	1.786	1.054
USAC-96	89.755	90.712	90.627	90.365	1.317	1.328	1.349	0.765
HE-AAC-32	44.406	57.396	58.264	53.355	2.543	2.655	2.719	1.594
HE-AAC-64	73.264	79.594	81.892	78.250	2.639	2.167	1.934	1.332
HE-AAC-96	88.821	89.774	89.321	89.305	1.265	1.370	1.496	0.793
AMR-32	57.557	43.618	52.717	51.297	3.026	2.357	2.819	1.641
3.5LP	21.236	20.590	20.208	20.678	1.162	1.197	1.061	0.656
7.0LP	41.873	39.009	38.330	39.737	1.999	1.885	1.759	1.088
Orig	99.802	99.755	99.759	99.772	0.125	0.144	0.161	0.083
VC-32	57.858	57.396	59.264	58.173	2.940	2.655	2.670	1.585
VC-64	73.264	79.594	81.892	78.250	2.639	2.167	1.934	1.332
VC-96	88.821	89.774	89.321	89.305	1.265	1.370	1.496	0.793
USAC-VC-32	12.160	2.557	4.108	6.275	2.829	1.710	1.729	1.280
USAC-VC-64	7.953	2.283	0.967	3.734	2.322	1.649	1.443	1.085
USAC-VC-96	0.934	0.939	1.307	1.060	1.199	1.095	1.523	0.738

1.2 Per Category and per Item Performance

This section presents a more detailed analysis of the test results. In Figures 7 to 17, the mapping between the labels (at right of figure) and the systems is as follows:

Label	System
Original	Original, uncoded item(s)
3.5 LP	Original signal low-pass filtered at 3.5 kHz
7.0 LP	Original signal low-pass filtered at 7.0 kHz
AMR-xy	AMR-WB+ at xy kbps
HE-AAC-xy	HE-AACv2 at xy kbps
USAC-xy	USAC at xy kbps
VC-xy	Virtual Codec at xy kbps (VC as defined in section 3)

Figures 7 to 9 show all average absolute scores in each test, including the scores for the low-pass anchors 3.5 LP and 7.0 LP (which correspond to the original signal low-pass filtered at, respectively, 3.5 and 7.0 kHz).

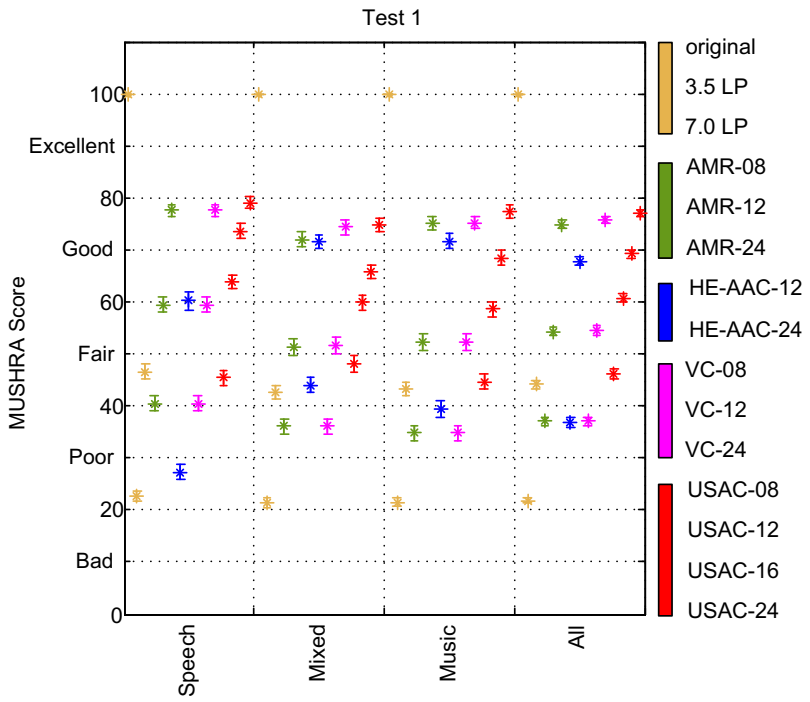


Figure 7. Average absolute scores in Test 1
per content type and over all items, with 95% confidence intervals.

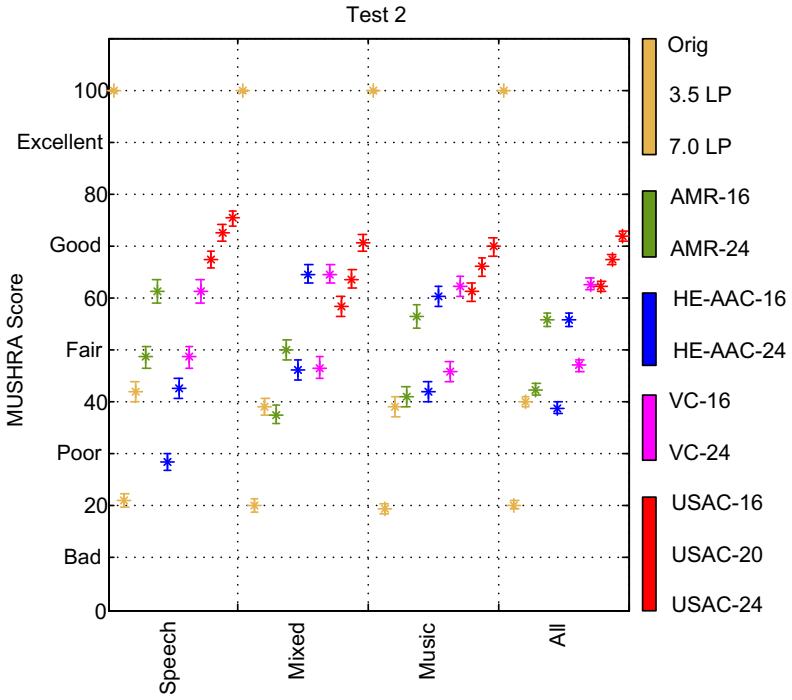


Figure 8. Average absolute scores in Test 2
per content type and over all items, with 95% confidence intervals.

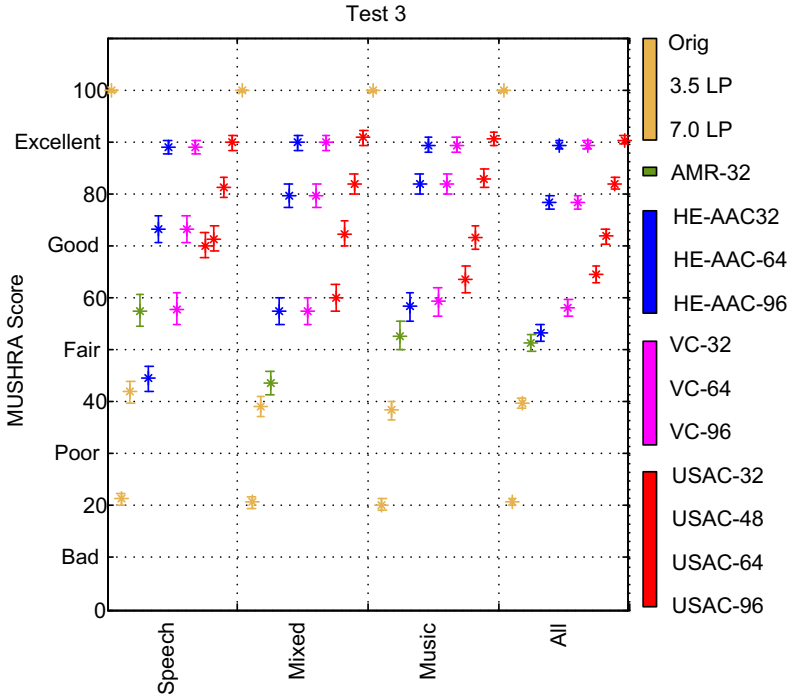


Figure 9. Average absolute scores in Test 3
per content type and over all items, with 95% confidence intervals.

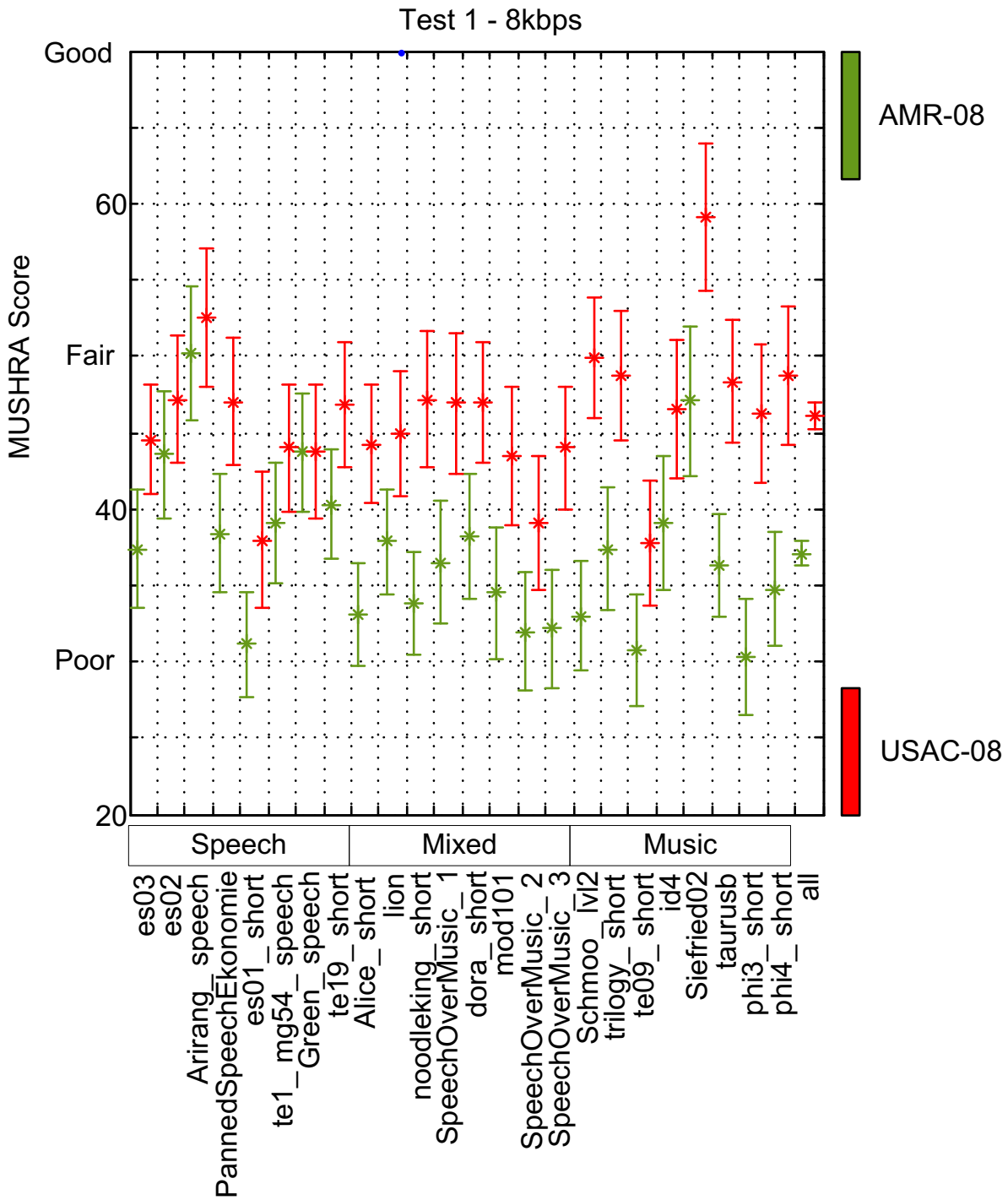


Figure 10. Average absolute scores per test item (category also shown) with 95% confidence intervals at 8 kbps mono.

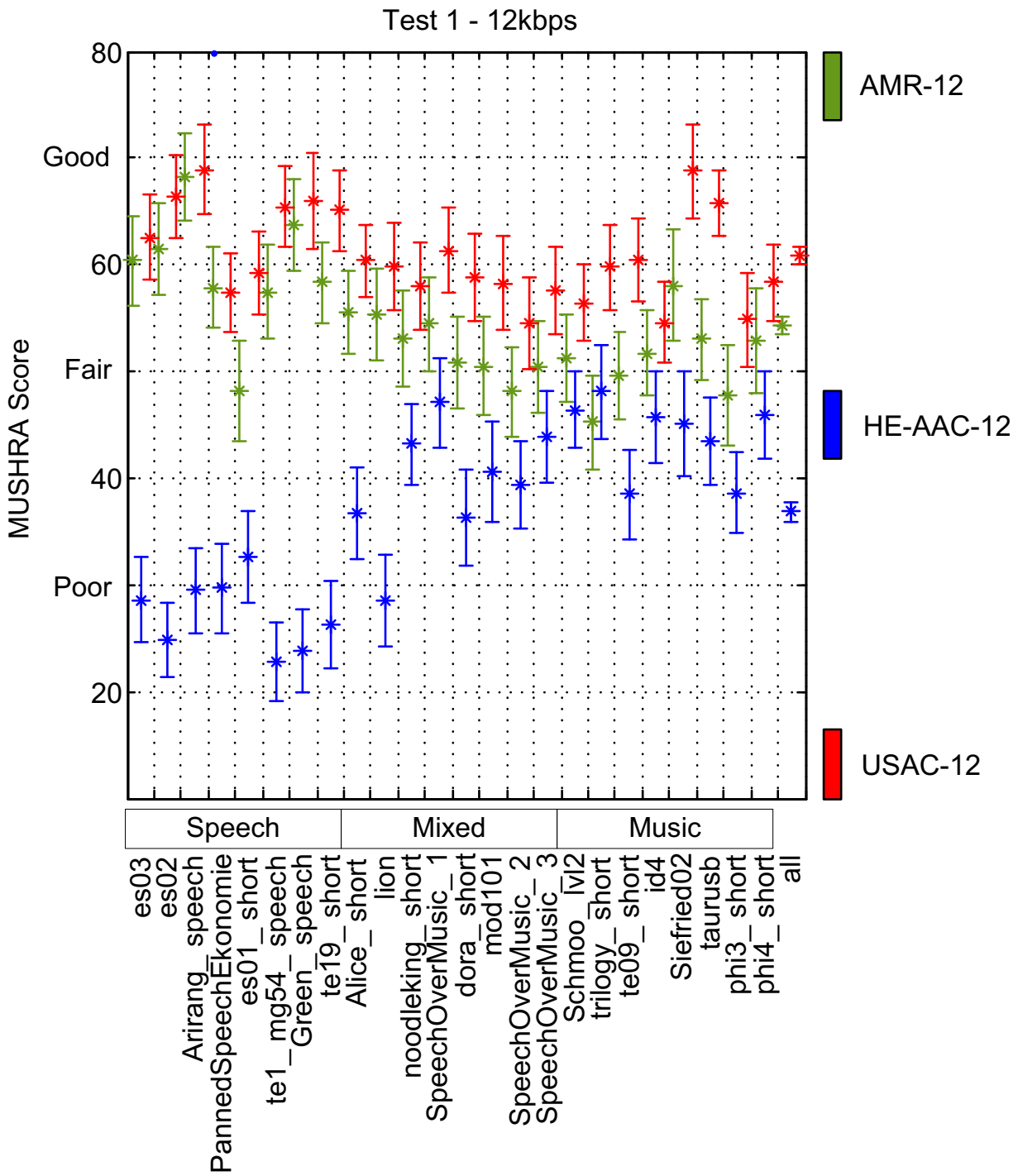


Figure 11. Average absolute scores per test item (category also shown) with 95% confidence intervals at 12 kbps mono.

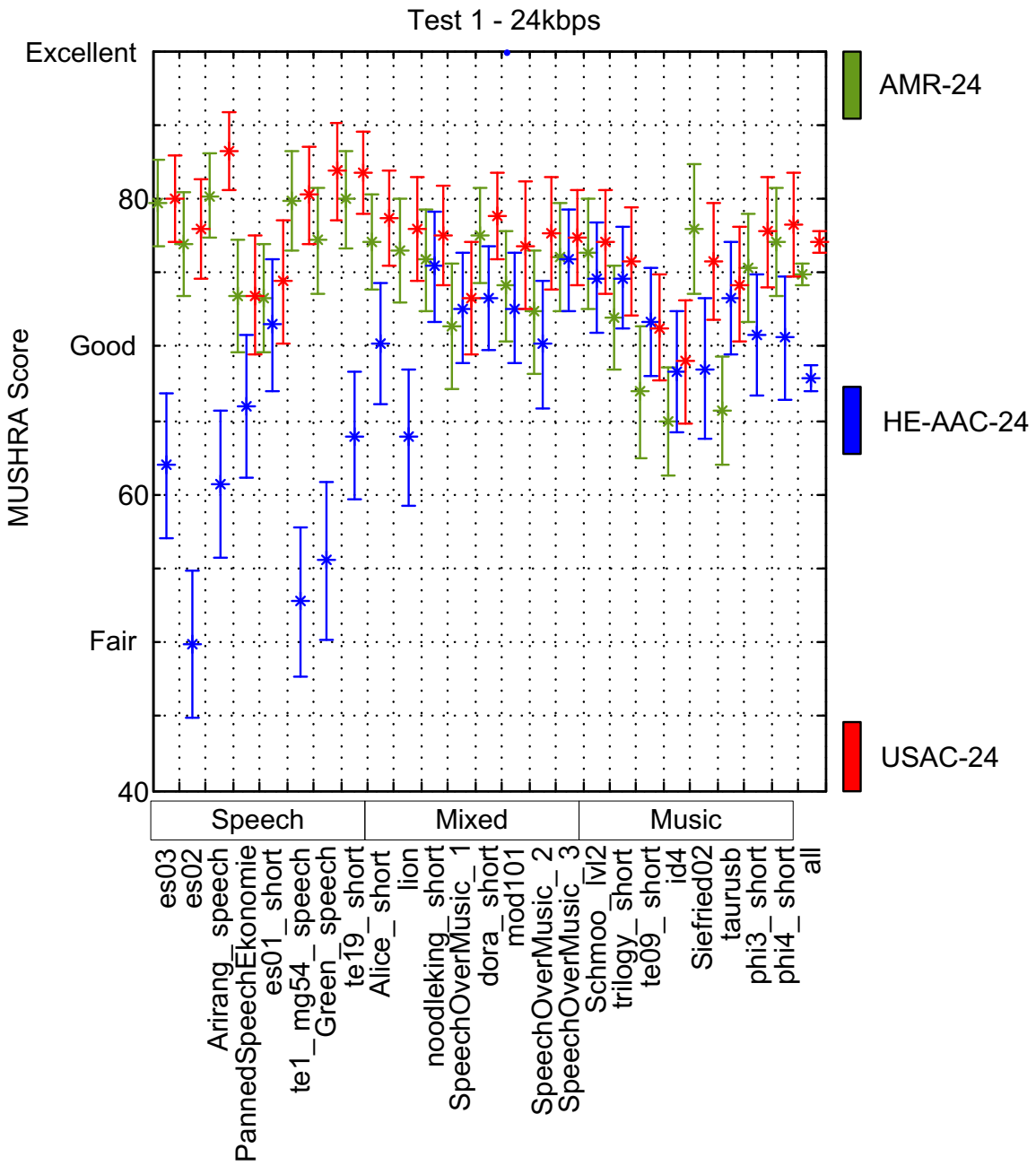


Figure 12. Average absolute scores per test item (category also shown) with 95% confidence intervals at 24 kbps mono.

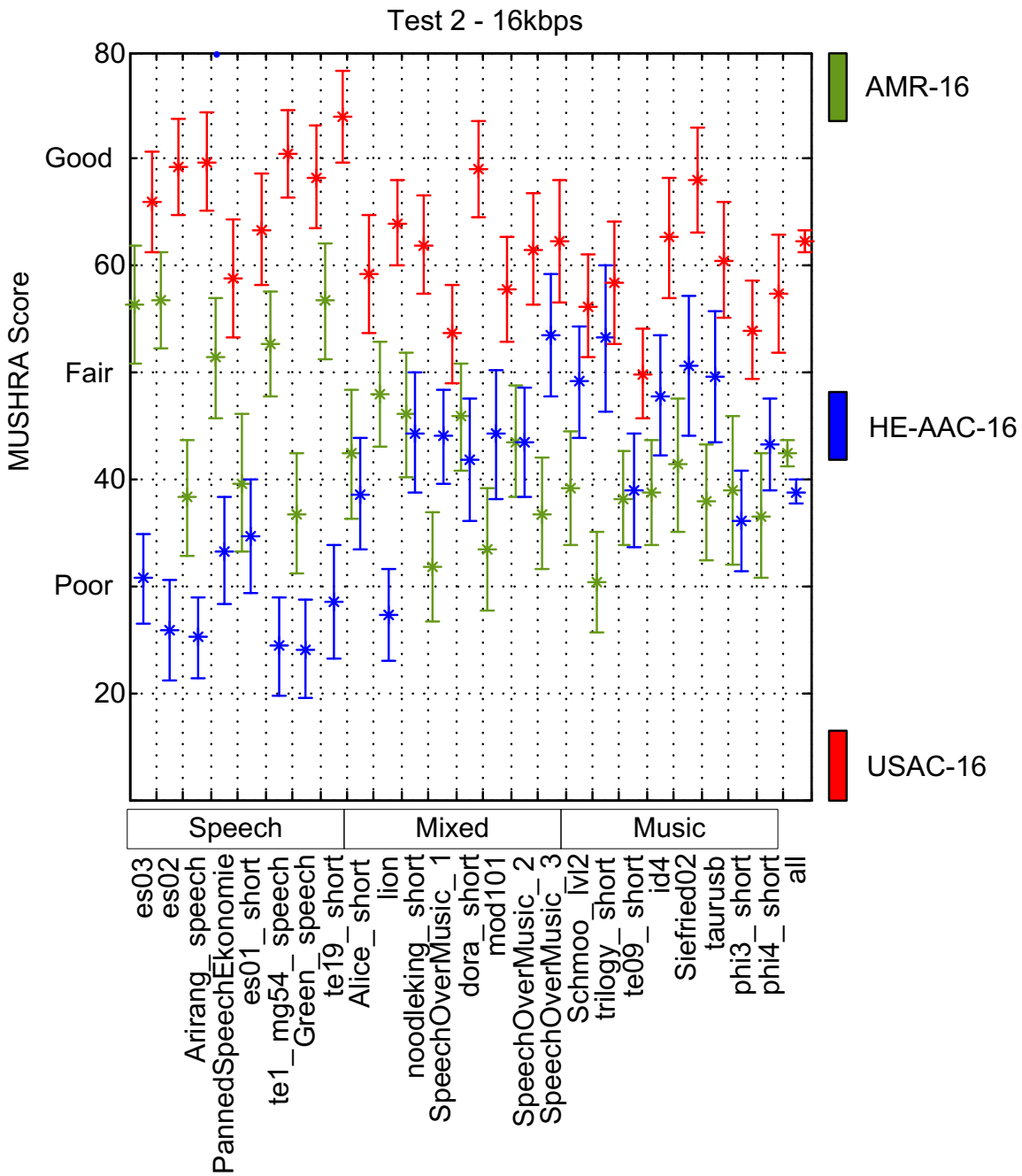


Figure 13. Average absolute scores per test item (category also shown) with 95% confidence intervals at 16 kbps stereo.

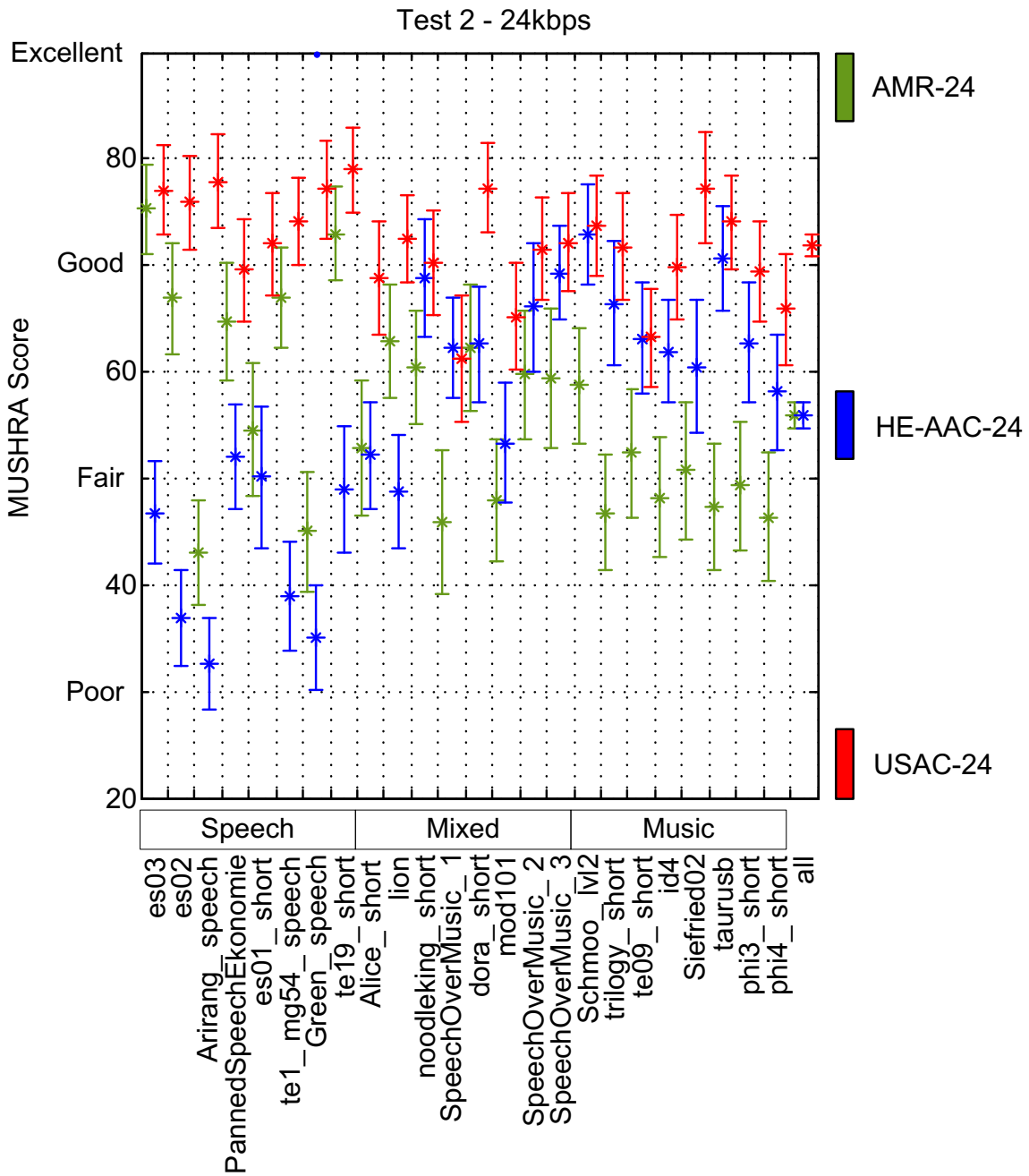


Figure 14. Average absolute scores per test item (category also shown) with 95% confidence intervals at 24 kbps stereo.

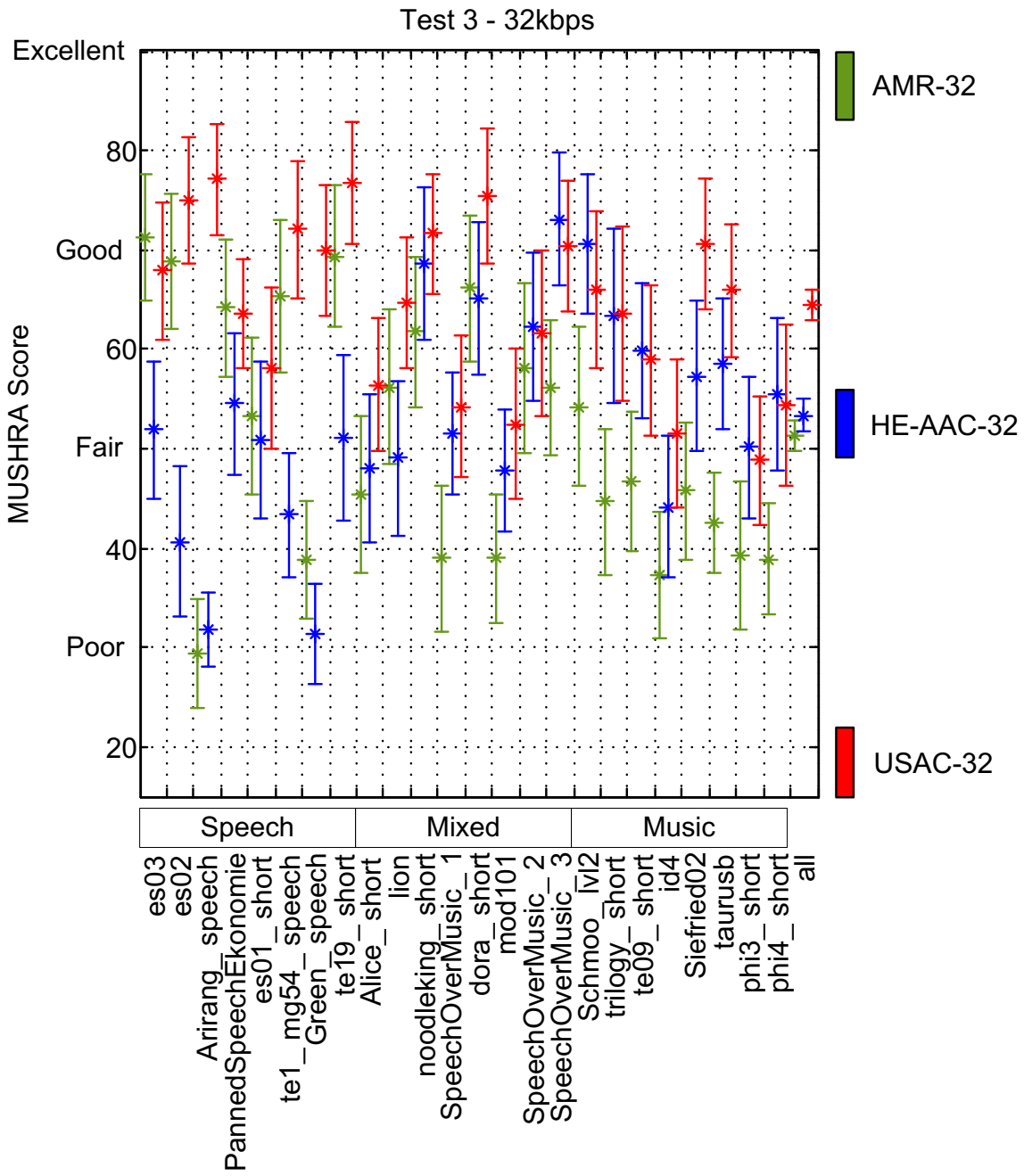


Figure 15. Average absolute scores per test item (category also shown) with 95% confidence intervals at 32 kbps stereo.

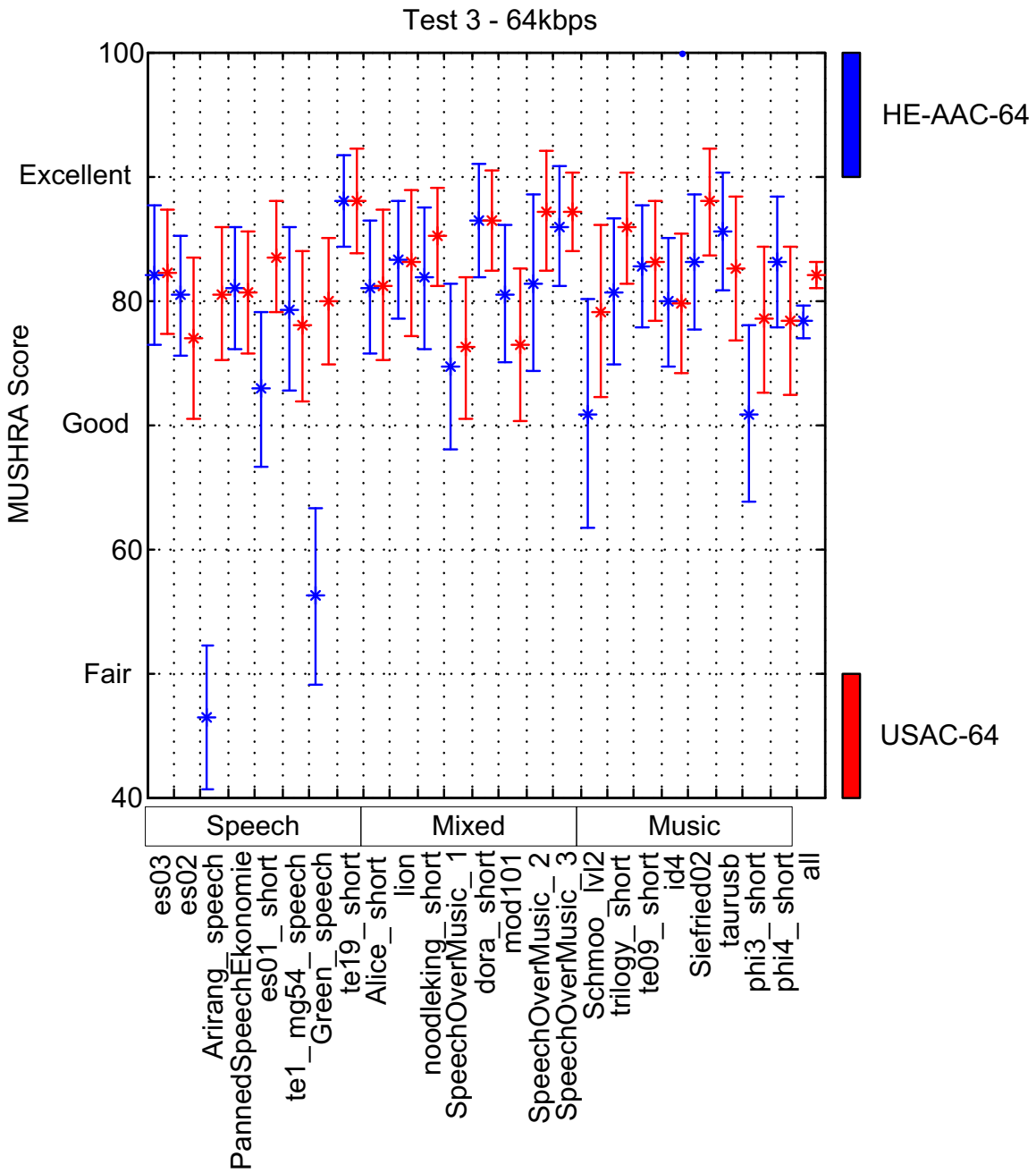


Figure 16. Average absolute scores per test item (category also shown) with 95% confidence intervals at 64 kbps stereo.

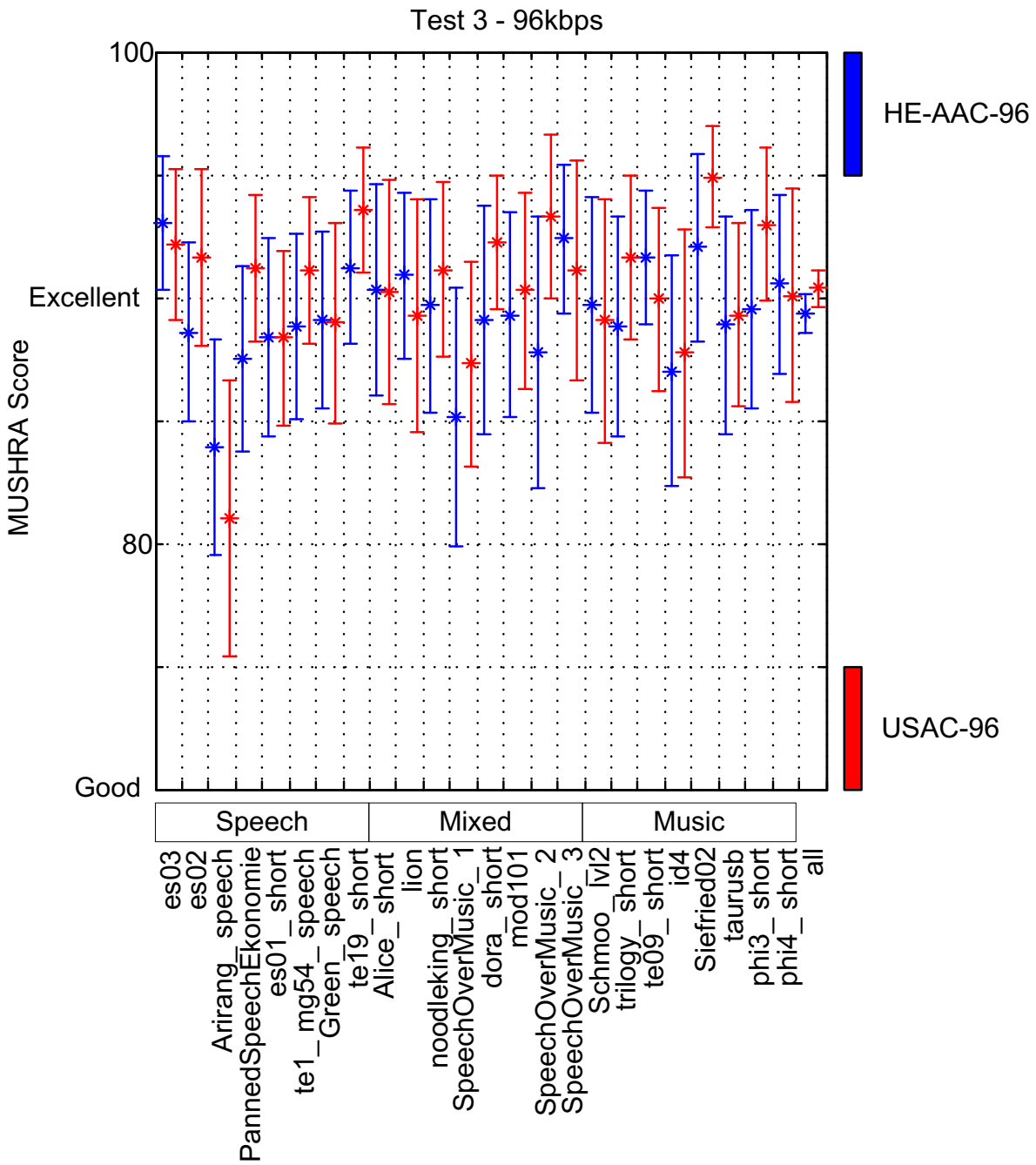


Figure 17. Average absolute scores per test item (category also shown) with 95% confidence intervals at 96 kbps stereo.

Annex 2 ANOVA Model for Subjective Data

1.1 Overview of ANOVA model

The objective of analysis of variance (ANOVA) is to assess whether a *treatment* applied to a set of samples has a significant effect, and to make that determination based on sound statistical principles [1], [2]. A treatment is, e.g., the processing of a signal by a coding system, but can also refer to other aspects of the experiment, so here we will to use the term *factor* instead of treatment.

The basic model of a score can be thought of as the sum of *effects*. A particular score may depend on which coding system was involved, which audio selection is being played, which laboratory is conducting the test, and which subject is listening. In other words, the score is the sum of a number of factor effects plus random error.

In terms of analyzing the data from the Verification Test, the following table lists the relevant factors in the experimental model. The test number (Test1, Test2, Test3) are not listed as factors since each test will be analyzed separately.

Factor	Description
Lab	Listening test site.
System	Coding system under test.
Category	The three signal content categories: Speech, Speech and Music or Music.
Signal	Speech or audio excerpt.

The factors System and Signal form a fully-balanced and randomized factorial design, in that in every Test all Signals were processed by all Systems and were presented to the listeners for grading in random order. This balance has the advantage that the mean score for each system is an appropriate statistic for estimating the quality of that system.

The factors System and Category are *fixed* in that they are specified in advance as opposed to being randomly drawn from some larger population.

The Signal factor is *nested* within the Category factor (see [1], Chapter 5) since each of the signals is classified into one of the categories (i.e. the 24 signals are organized as distinct groups of 8 signals each of which is assigned to, or nested within, one of the 3 categories).

Signal would be a *random* factor if the signals were actually selected at random from the population of all possible signals. Intuitively this is very appealing in that we might want to know how well the coding systems will perform for all possible speech and music items. However, we want the best coding system so the speech and music items were specifically selected because they are “difficult” items to code and so represent the “right tail” of a distribution of items rather than the entire population. Hence we have chosen to model Signal as a fixed factor.

The Labs, or test sites, was modeled as a random factor in that each Lab represent a specific test apparatus (i.e. listening room and audio presentation system) from a universe of possible test sites.

Since each Lab has a distinct set of listeners, the Listener factor is nested within the Labs factor. Listeners could be viewed as a random factor, in that it is intuitive and appealing to consider the listeners that participated in the test as representative of the larger population of all listeners. In this case the test outcome would represent the quality that would be perceived by the “typical” listener. However, the goal of the test was to have maximum discriminating capability so as to identify the best performing system. To this end, the subjects used were very experienced listeners that were “experts” at discerning the types of distortion typical of low-rate speech and audio coding. Regardless of these considerations, Listener was not used as a factor because of the very large number of levels (i.e. distinct listeners).

Two aspects of the experimental design were not optimal. First, the Lab and Listener factors are not balanced. Participation as a test site and as a listener was voluntary, and a balanced design would have all sites and all listeners scoring all Tests, Systems and Signals, which was beyond the resources available

within the MPEG Audio subgroup. However, the ANOVA calculations take the imbalance into account when computing the effects of each factor. Second, each test was split into two subtests. Each subtest was balanced in that all listeners rated every system for every signal, however at any given test site typically different listeners participated in each subtest.

An important issue in using ANOVA is that it relies on several assumptions concerning the data set and the appropriateness of these assumptions should be checked as part of the data analysis. The most important assumptions are:

- The error has a Gaussian distribution.
- The variance of the error across factor levels is constant.

Finally, note that all ANOVA calculations, histogram and standard probability plots were performed using the R statistical package [3], [4].

1.2 Test 1

Model

An aspect of ANOVA is to test the suitability of the model. The simplest model incorporating all factors is expressed as:

```
Score = Lab + System + Signal + Signal in Category + Eror.
```

The last element in the model "Signal in Category" indicates that Signal is nested within Category.

The ANOVA report when using this model is:

```
               Df  Sum Sq Mean Sq  F value    Pr(>F)
lab            12  835627   69636  376.817 < 2.2e-16 ***
sys            11 7638267  694388 3757.524 < 2.2e-16 ***
sig:cat        23   53247    2315   12.527 < 2.2e-16 ***
Residuals   17664 3264296     185
```

The report indicates that all model factors are highly significant.

A more complex model captures interaction between System and Signal, in that the performance of a coding system might depend on the specific signal being coded. Finally, we anticipate that there may also be an interaction between the coding systems and the testing laboratories. The statistical significance and the size of the effects due to factors and interactions will be a measure of the importance of the factor and interaction terms in the model.

The following model incorporates all factors and interactions that turned out to be significant:

```
Score = Lab + System + Category
          + Signal in Category
          + Category and System interaction
          + Lab and Signal interaction
          + Lab and System interaction
          + Signal and System interaction
          + Error.
```

The ANOVA report when using this model is:

```
               Df  Sum Sq Mean Sq   F value    Pr(>F)
lab            12  835627   69636  491.1320 < 2.2e-16 ***
sys            11 7638267  694388 4897.4414 < 2.2e-16 ***
sig:cat        23   53247    2315   16.3279 < 2.2e-16 ***
cat:sys        22  187065    8503   59.9703 < 2.2e-16 ***
lab:sig       276  171987     623    4.3949 < 2.2e-16 ***
```

```
lab:sys       132  336048    2546   17.9554 < 2.2e-16 ***
sig:sys       231  158411     686    4.8366 < 2.2e-16 ***
Residuals   17003 2410785     142
```

The report indicates that all model factors and interactions are highly significant.

Performance

Using an ANOVA model does not change the mean score of the system under test. However, because it removes the factor mean effects from the error term, it reduces the error variance and hence the confidence interval on the mean scores. The CI Value (i.e. the ± value used to compute the 95% confidence interval) is computed as:

CI Value = qt(0.975, degree_freedom) * sqrt(residual_mse/n_obs)

Where **qt()** is the cumulative Student-t distribution. For Test1 this is

CI Value = qt(0.975, 12684) * sqrt((2058031/12684)/1476)

The following is a table of mean scores and CI Value for Test 2 for the systems under test when averaged across all Labs and Signals and with listener post screening applied.

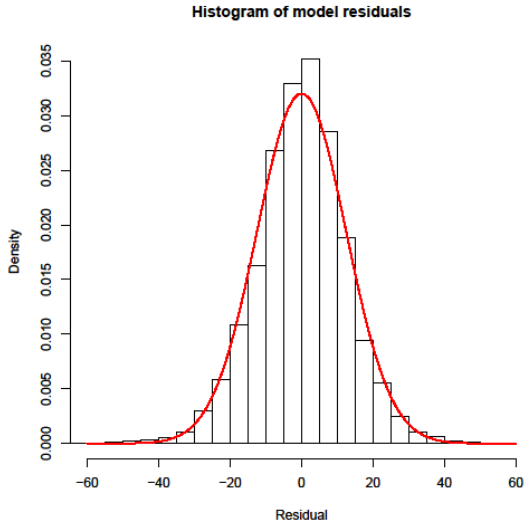
Sys	Mean	CI Value
usac_08m	46.012	0.608
usac_12m	60.764	0.608
usac_16m	69.268	0.608
usac_24m	77.056	0.608
heaacv2_12m	36.820	0.608
heaacv2_24m	67.771	0.608
amrwbp_08m	37.038	0.608
amrwbp_12m	54.276	0.608
amrwbp_24m	74.858	0.608
lp35	21.849	0.608
lp70	44.138	0.608
orig	99.915	0.608

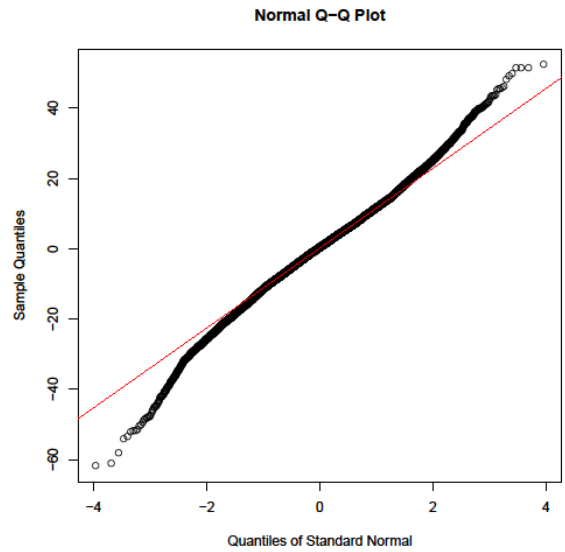
Verification of model assumptions

The following plots verify that the ANOVA residual has approximately a Gaussian distribution, as required for the validity of the ANOVA. Note that the systems Hidden Reference, 7.0 kHz low-pass original and 3.5 kHz low-pass original are removed prior to testing the ANOVA model assumptions since these systems do not get a truly random subjective assessment: subjects are instructed to score the Hidden Reference at 100 and generally tend to score the 7.0 kHz low-pass original and 3.5 kHz low-pass original as some nearly fixed score whose value is based on personal preference.

The left-hand plot below shows a histogram of the Test1 residual with a best-fit Gaussian distribution (shown in red) superimposed on top. The right-hand plot shows a Normal Q-Q Plot of a Gaussian distribution (red line) and the Test1 residuals. The plot is such that a true Gaussian distribution lies on a straight line. One can see that the Test1 residual deviates from the red line only at the ends, i.e. the tails of the distribution.

Both plots suggest that distribution of the Test1 residuals is a very close to a Gaussian.

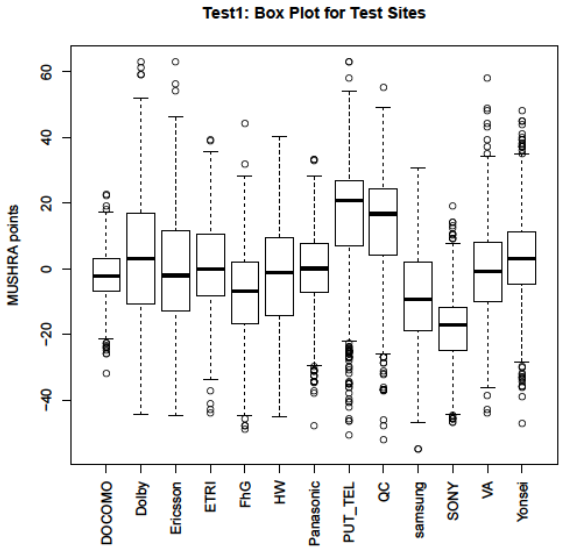


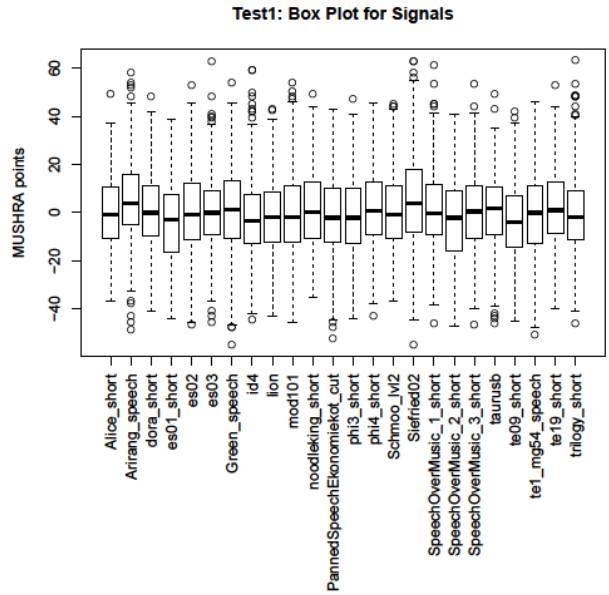


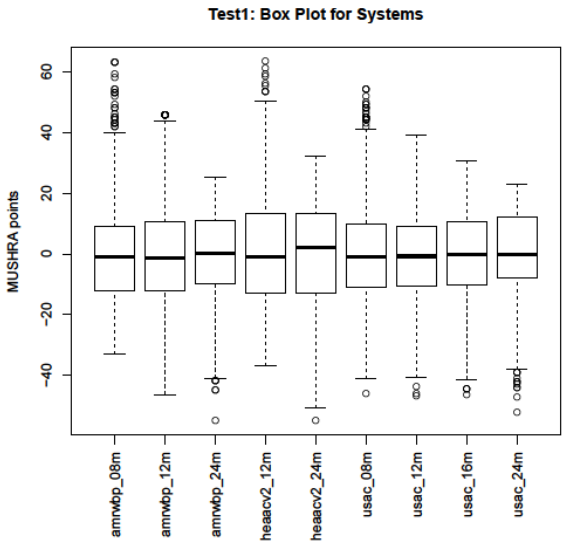
The following box plots show the scores associated with each level of the factors (or factor values). For each of the factors Lab (Test Site),, Sample and System, the box plots indicate the distribution of score values after the factor effect is removed. In the box plots:

- The box indicates the range of the middle two quartiles of data (i.e. the box encompasses ±25% of the data, as measured from the mean).
- The "whiskers" indicate ±37.5% of the data, as measured from the mean
- The "circles" indicate data outliers that lie beyond of the ±37.5% region.

The plots indicate that the residuals have the approximately the same distribution: Test Site spread is within a factor of 2 to 3, while Signal and System is within a few tens of percent.







1.3 Test2

Model

The following model incorporates all factors and interactions that turned out to be significant:

```
Score = Lab + System + Category
          + Signal in Category
          + Category and System interaction
          + Lab and Signal interaction
          + Lab and System interaction
          + Signal and System interaction
          + Error.
```

The ANOVA report when using this model is:

```
              Df  Sum Sq Mean Sq   F value    Pr(>F)
lab            7  303732   43390  273.3815 < 2.2e-16 ***
sys            9 4082903  453656 2858.2672 < 2.2e-16 ***
sig:cat       23   54898    2387   15.0384 < 2.2e-16 ***
cat:sys       18  214642   11925   75.1308 < 2.2e-16 ***
lab:sig      161   57296     356    2.2422 < 2.2e-16 ***
lab:sys       63  193316    3069   19.3332 < 2.2e-16 ***
sig:sys      189  171833     909    5.7283 < 2.2e-16 ***
Residuals   8889 1410836     159
```

The report indicates that all model factors and interactions are highly significant.

Performance

The following is a table of mean scores and CI Value for Test 2 for the systems under test when averaged across all Labs and Signals and with listener post screening applied.

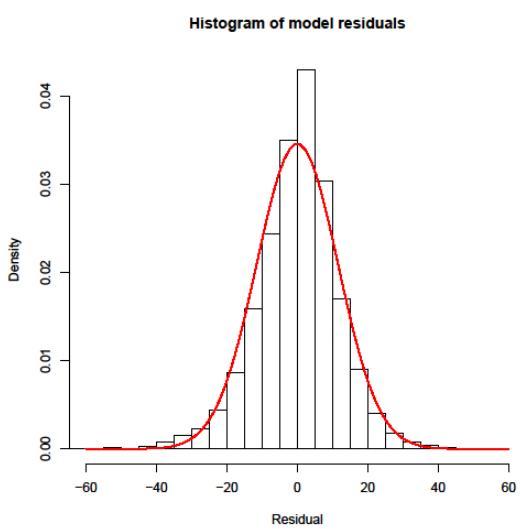
Sys	Mean	CI
usac_16s	62.269	0.807
usac_20s	67.423	0.807
usac_24s	71.915	0.807
heaacv2_16s	38.733	0.807
heaacv2_24s	55.848	0.807
amrwbp_16s	42.384	0.807
amrwbp_24s	55.886	0.807
lp35	20.280	0.807
lp70	40.088	0.807
orig	99.921	0.807

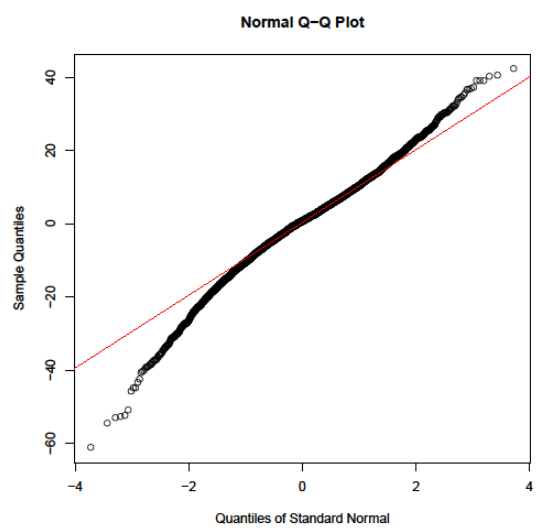
Verification of Model Assumptions

The following plots verify that the ANOVA residual has approximately a Gaussian distribution. As in Test 1, the systems Hidden Reference, 7.0 kHz low-pass original and 3.5 kHz low-pass original are removed prior to testing the ANOVA model.

The left-hand plot below shows a histogram of the Test2 residual with a best-fit Gaussian distribution (shown in red) superimposed on top. The right-hand plot shows a Normal Q-Q Plot of a Gaussian distribution (red line) and the Test2 residuals.

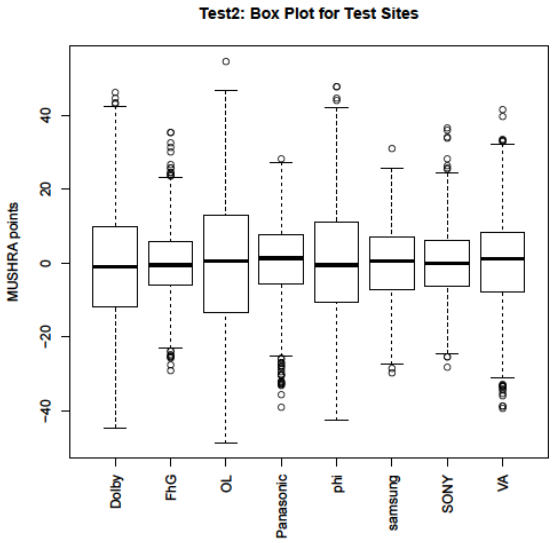
Both plots suggest that distribution of the Test1 residuals is a very close to a Gaussian.

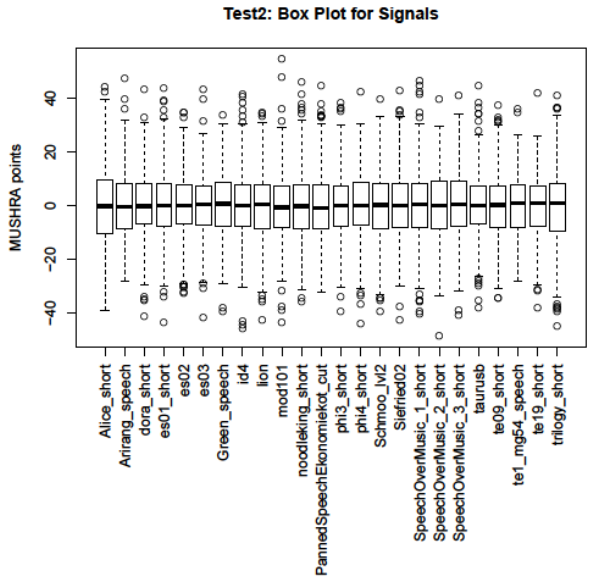


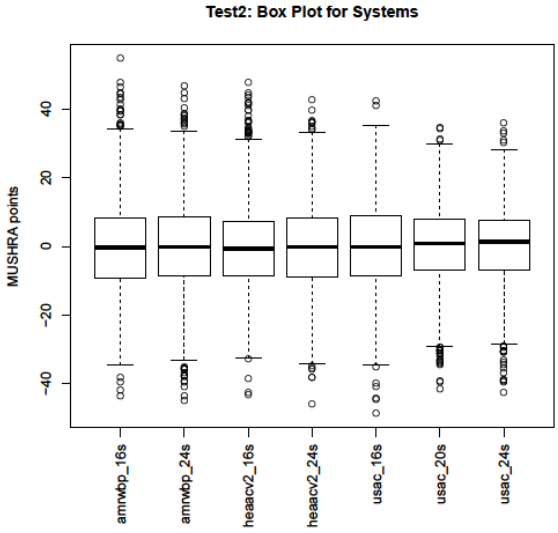


The following box plots show the scores associated with each level of the factors (or factor value) Lab (Test Site), Signal and System.

The plots indicate that the residuals have the approximately the same distribution: Test Site spread is within a factor of 2 to 3, while Signal and System is within a few tens of percent.







1.4 Test3

Model

The following model incorporates all factors and interactions that turned out to be significant:

```
Score = Lab + System + Category
          + Signal in Category
          + Category and System interaction
          + Lab and Signal interaction
          + Lab and System interaction
          + Signal and System interaction
          + Error.
```

The ANOVA report when using this model is:

```
              Df  Sum Sq Mean Sq   F value    Pr(>F)
lab            5  309433   61887  492.4650 < 2.2e-16 ***
sys           10 3700179  370018 2944.4358 < 2.2e-16 ***
sig:cat       23   75859    3298   26.2459 < 2.2e-16 ***
cat:sys       20   66511    3326   26.4631 < 2.2e-16 ***
lab:sig      115   34466     300    2.3849 1.291e-14 ***
lab:sys       50  101405    2028   16.1388 < 2.2e-16 ***
sig:sys      210  160807     766    6.0935 < 2.2e-16 ***
Residuals   6562  824626     126
```

The report indicates that all model factors and interactions are highly significant.

Performance

The following is a table of mean scores and CI Value for Test 3 for the systems under test when averaged across all Labs and Signals and with listener post screening applied.

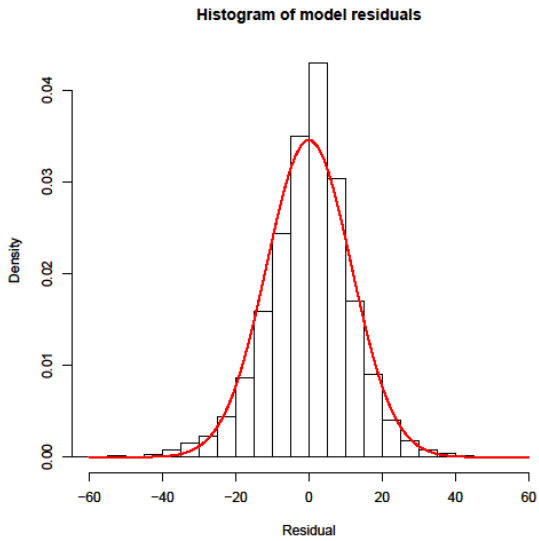
Sys	Mean	CI Value
usac_32s	64.448	0.871
usac_48s	71.744	0.871
usac_64s	81.984	0.871
usac_96s	90.365	0.871
heaacv2_32s	53.355	0.871
heaacv2_64s	78.250	0.871
heaacv2_96s	89.305	0.871
amrwbp_32s	51.297	0.871
lp35	20.678	0.871
lp70	39.737	0.871
orig	99.772	0.871

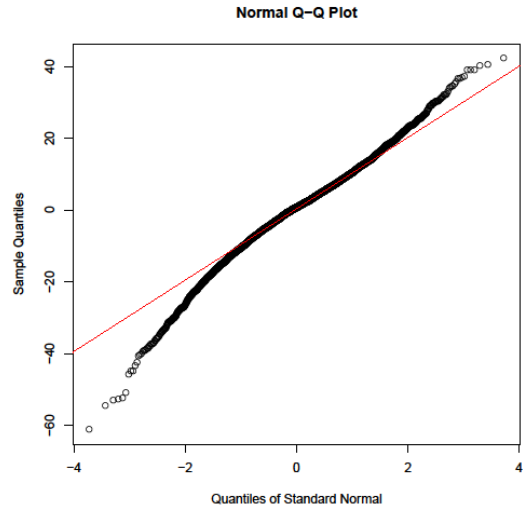
Verification of Model Assumptions

The following plots verify that the ANOVA residual has approximately a Gaussian distribution, as required for the validity of the ANOVA. As in Test 1, the systems Hidden Reference, 7.0 kHz low-pass original and 3.5 kHz low-pass original are removed prior to testing the ANOVA model.

The left-hand plot below shows a histogram of the Test3 residual with a best-fit Gaussian distribution (shown in red) superimposed on top. The right-hand plot shows a Normal Q-Q Plot of a Gaussian distribution (red line) and the Test3 residuals.

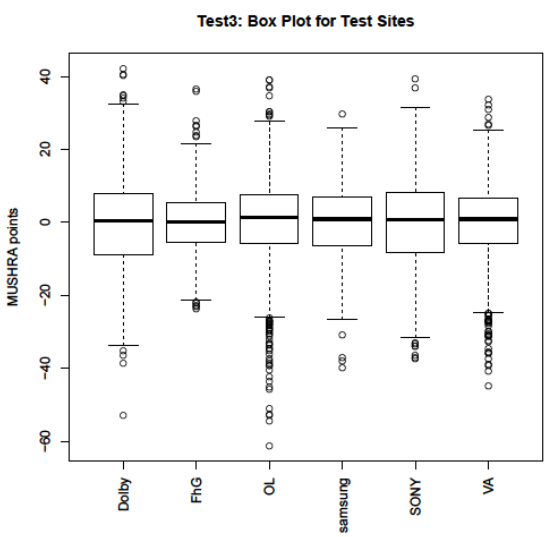
Both plots suggest that distribution of the Test1 residuals is a very close to a Gaussian.

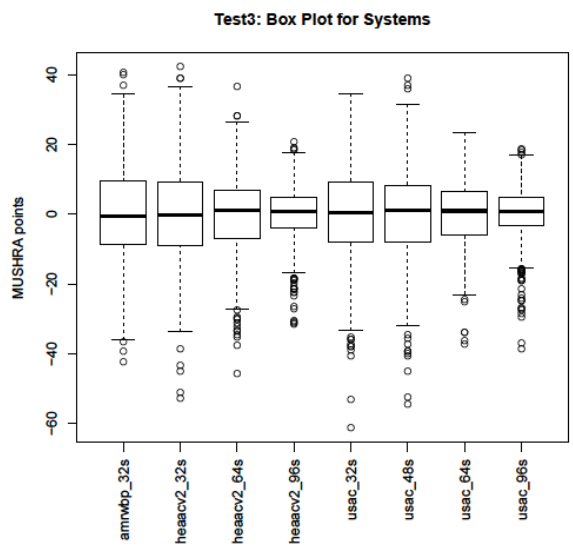


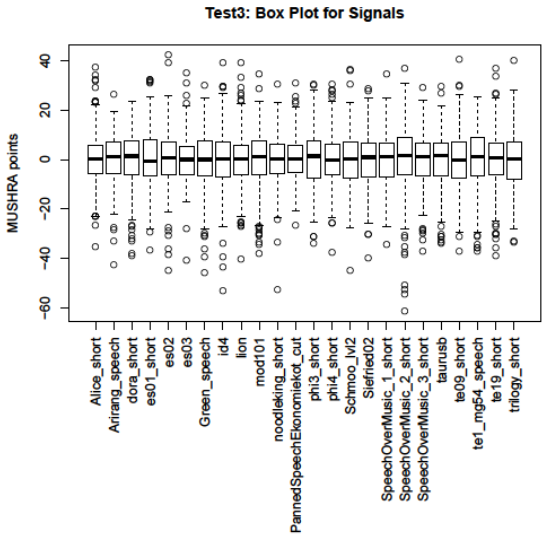


The following box plots show the scores associated with each level of the factors (or factor value) Lab (Test Site), Signal and System.

The plots indicates that the residuals have the approximately the same distribution: Test Site spread is within a factor of 2, while Signal and System is within a few tens of percent.






Test3: Box Plot for Signals
MUSHRA points
40
20
0
−20
−40
−60
Alice_short
Arirang_speech
dora_short
es01_short
es02
es03
Green_speech
id4
lion
mod101
noodleking_short
PannedSpeechEkonomiekot_cut
phi3_short
phi4_short
Schmoo_lvl2
Siefried02
SpeechOverMusic_1_short
SpeechOverMusic_2_short
SpeechOverMusic_3_short
taurusb
te09_short
te1_mg54_speech
te19_short
trilogy_short

1.5 Annex 2 References

[1] Montgomery, D.C. *Design and Analysis of Experiments*. John Wiley and Sons, New York, 1976.

[2] Bech, S. and Sacharov, N. *Perceptual Audio Evaluation, Theory, Method and Application*. John Wiley and Sons, Chinchester, West Sussex, England, 2002.

[3] Venables, W. N. and Ripley, D. B. *Modern Applied Statistics with S, Fourth Edition*. Springer, New York, 2002.

[4] The R Project for Statistical Computing, http://www.r-project.org/

ANNEX 3 -- Listener Instructions

Listeners must read these instructions and participate in the indicated training phase prior to their participation in the test phase.

Listener Instructions
for
USAC Verification test

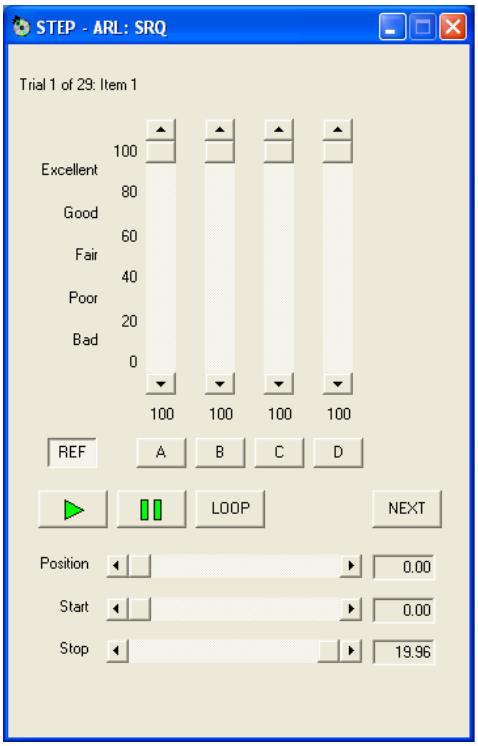
1.1 Introduction

The MPEG Audio Subgroup is undertaking work to verify the subjective quality of the Unified Speech and Audio Coding (USAC) standard in various operating modes.

This test will use the MUSHRA test methodology, which has the advantage of displaying all stimuli (both coding systems and anchor systems) for a given test item. Hence you are able to directly compare the stimuli in the course of giving a grade to each. Test administrators are encouraged to read the MUSHRA specification [3] to obtain an understanding of the test procedure and test objectives.

The Figure below shows the user-interface presenting one test item as processed by all systems under test. The buttons represent the reference (REF), which is always displayed at the bottom left, and all the systems under test, including the codecs under test, reference codecs, hidden reference and anchor signals (band-limited processed references), which are displayed as letter buttons. Above each button, with the exception of the button for the reference, a slider permits the listener to grade the quality of the test item on a continuous quality scale. The descriptors associated with the scale are “bad” (0-20), “poor” (20-40), “fair” (40-60), “good” (60-80) and “excellent” (80-100). For each of the test items, the systems under test are randomly assigned to the buttons. In addition, the order of presenting the test items is randomized for each listener.

Note that the audio presentation switches instantly and seamlessly from one system to another. Clicking on the “Loop” button plays the signal continuously. The horizontal Position slider indicates the instantaneous position in the signal waveform. Grabbing and moving the Start slider alters the start point for waveform looping, and similarly moving the Stop slider alters the end point, thus permitting a “loop and zoom” function that is particularly powerful for subjective evaluation. Rate the processed signals by grabbing and moving the vertical sliders above their corresponding letter buttons. When you are satisfied with the ratings, click on the “Next” button to go on to the next trial. When the last trial is scored, the MUSHA window is replaced by the Administrator window. Notify the test administrator that you have completed the listening session.



1.6 Training phase

The training phase will be conducted individually. You will be participating in a simulated test whose graphical interface is exactly like the actual test, but the training is both shorter than the actual test and uses test items that are different from those in the actual test. Please listen to the training signals to understand how processed signals sound relative to the reference.

In the course of listening to these signals, please:

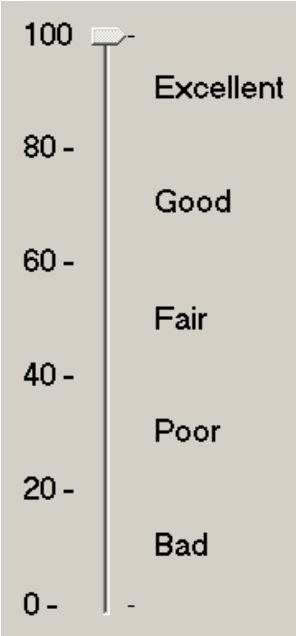
- Listen to a test item several times.
- Rate overall sound quality, not merely the quality of a portion of a test item, for example, only the portion in which speech is active or in which music is active.
- Note that if two systems have a different perceived quality with respect to the Reference they should receive a different score.

Bear in mind that any change from the original is to be considered an impairment.

1.7 Test phase

The test phase will be carried out individually in test sessions each lasting about 20 to 30 minutes. In each trial, you will hear several versions of the test items, each processed by a different system under test. One of the "systems under test" is actually the Reference. You are asked to identify that system by giving it a score of 100. Note that it is perfectly acceptable to give more than one system a score of 100.

You are asked to judge the "Overall Sound Quality" of the versions of the test item in each trial. This attribute is related to any and all differences between the reference and the coded/decoded test item, including speech quality, music quality, background noise, audio distortion, signal bandlimiting and stereo image, such that any difference between the reference and the coded/decoded test item is to be considered an impairment. The assessment is to be done on a scale from 0 to 100, as shown here:


100
Excellent
80 -
Good
60 -
Fair
40 -
Poor
20 -
Bad
0 -