INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC 1/SC 29/WG 11 N7137

April 2005, Busan, Korea

Source: Audio Subgroup
Title: Listening test report on MPEG-4 High Efficiency AAC v2
Status: Approved

Summary

This document reports on a verification test of the MPEG-4 Parametric Stereo (PS) coding tool which, when combined with the MPEG-4 Advanced Audio Coding Low Complexity (AAC LC) coding tool and the MPEG-4 Spectral Band Replication (SBR) coding tool, comprises the MPEG-4 High Efficiency AAC v2 (HE AAC v2) Profile.

The verification test compares the performance of the MPEG-4 High Efficiency AAC Profile v2 coder to that of the MPEG-4 High Efficiency AAC (HE AAC) Profile coder (i.e. AAC LC tool in combination with the SBR tool). The verification test shows that the HE AAC v2 codec, in mean performance, offers a coding gain of 25% as compared to the performance of HE AAC.

Table of Contents

1 Introduction

1.1 Background

In mid-1999 the International Standard ISO/IEC 14496-3, MPEG-4 Audio Version 1 issued and in early 2000 the ISO/IEC 14496-3 / AMD1, MPEG-4 Audio Version 2 issued. Extensive tests have been conducted by MPEG [3,4] to verify that the MPEG-4 standard contains state of the art technology. However, WG11 is always interested in new developments which may provide improvements over the existing MPEG-4 standard and which may lead to extensions of MPEG-4 or to new work items. For this reason, at the 53rd MPEG meeting, in Beijing, MPEG issued a Call for Evidence Justifying the Testing of Audio Coding Technology [5]. Evidence submitted in response to the Call was examined at the 55th MPEG meeting, in Pisa, and it was determined that there was technology that might improve upon the MPEG-4 standard. Based on the results of the Call for Evidence, work was begun in WG11 to standardize technology for an MPEG-4 Bandwidth Extension tool that could be applied to general audio signals, and a Parametric Audio Coder both aiming at higher quality than existing audio coders in MPEG [2].

The work on a bandwidth extension tool led to the standardisation of SBR (Spectral Band Replication) as a tool that could be combined with MPEG-4 AAC. The MPEG-4 High Efficiency AAC (HE AAC) Profile incorporates both the SBR tool and the Low Complexity AAC (AAC) tool.

The work on the Parametric Coder, led to the development of a Parametric Stereo tool. It was shown that this tool could advantageously be combined with the HE AAC Profile decoder, and thus the combination of the HE AAC profile decoder and the Parametric stereo tool was defined as the High Efficiency AAC v2 Profile.

1.2 The High efficiency AAC v2 Profile codec

The two extensions to the MPEG-4 AAC LC decoder (the AAC Profile decoder) form fully backwards compatible decoders as is outlined in Figure 1.

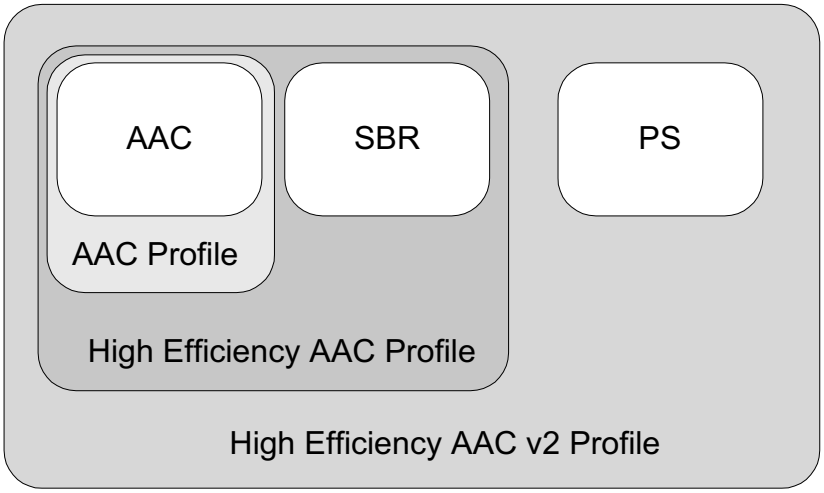

Figure 1 - The modular approach of the AAC, HE AAC and HE AAC v2 Profile decoders

Due to the modular approach of the HE AAC and HE AAC v2 codecs, an HE AAC v2 encoder does not necessarily need to use the PS tool when encoding stereo and can thus produce fully backwards compatible stereo HE AAC bitstreams (or even AAC bitstreams without SBR). A HE AAC v2 encoder will make use of the SBR and PS tool as appropriate for the required compression performance.

A further consequence of the modular approach is decoder “forward” compatibility: an AAC decoder is able to decode the AAC part of an HE AAC v2 bitstream which contains both, SBR and PS data. However playback quality will be significantly limited. It is therefore recommended to use this kind of compatibility only in cases where it makes commercial sense. For the mobile environment, it is highly recommended to use the HE AAC

v2 codec, as it provides the highest compression efficiency at low data rates, while still being MPEG compliant and compatible with all higher bitrate variants up to transparent AAC coding.

Table 1 illustrates the bitstream and decoding compatibilities as outlined above.

Table 1 - Compatibility between the AAC, HE AAC and HE AAC v2 Profile decoders

		Decoder		
		HE AAC v2	HE AAC	AAC
Encoder mode	HE AAC v2	Yes	Mono only	Mono only, no SBR
	HE AAC	Yes	Yes	No SBR
	AAC	Yes	Yes	Yes

1.3 Test Methodology

For the verification of the parametric stereo technology, a MUSHRA test (see appendix A.1) was performed. The MUSHRA test compared the performance of MPEG-4 HE AAC v2 with that of MPEG-4 HE AAC.

2 Codecs under test

There were two codecs under test. The MPEG-4 High Efficiency AAC Profile codec, which was used as a reference of the current state of the art MPEG-4 compression technology, and the High Efficiency AAC v2 Profile codec. As specified in the MUSHRA test methodology, a hidden reference and two band-limited versions of the reference were included as anchors and references in the tests. The codecs under test are shown in Table 2, which also shows the labels used for each codec in the tables and plots throughout the remainder of this report.

Table 2 - Codecs Under Test

Coding Scheme	Label	Bit rate
MPEG - High Efficiency AAC	HE-AAC 24	24 kbps stereo
	HE-AAC 32	32 kbps stereo
MPEG - 4 High Efficiency AAC v2	HE-AAC v2 24	24 kbps stereo
Anchors and References	H-Ref-Orig	16-bit PCM, mono
	H-Ref-3.5	16-bit PCM, mono
	H-Ref-7	16-bit PCM, mono

3 Test material

The items used for the test were the same as used for the formal verification test of the HE AAC profile codec [6]. They were selected from 50 potential candidates, by a selection panel at France Télécom R&D. The ten items listed in Table 3 were used in the test.

Table 3 - 10 Selected items for the MUSHRA test.

Item No.	filename	signal
1	te01	Dorita
2	te04	Harpsichord
3	te07	Male German Speech
4	te09	Tracy Chapman
5	te16	Accordion/Triangle
6	te20	George Duke
7	te33	<CROISEMENT I> pour hautbois, violon et contrebasse
8	te41	fanfare
9	te44	Bransle
10	te48	Layla

4 Test Centers

The tests took place at Philips and Coding Technologies. In total 18 expert listeners participated. A computer based MUSHRA presentation was used, and the playback devices were STAX Lambda Pro open headphones. There was only one listener at the time in the listening room due to the open headphones.

5 Test Results

A statistical analysis and post screening (see appendix A.2) was done on the listening test data. Figure 2 display the mean values (horizontal tick) and 95% confidence intervals (vertical tick) for every items, and over all items for every coding scheme.

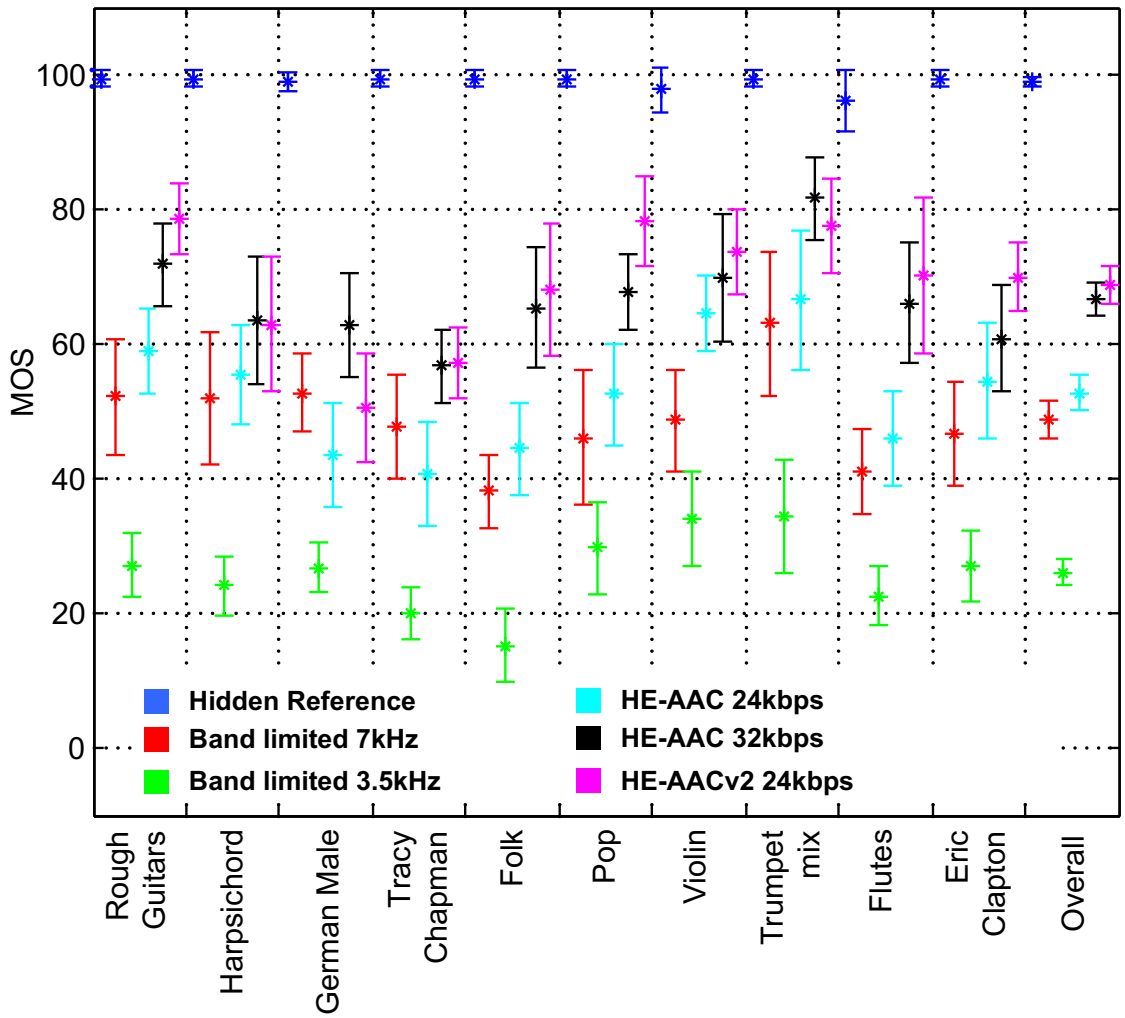

Figure 2 - Test results per item and overall

Comparing the MPEG-4 HE AAC v2 codec operating at 24kbps with MPEG-4 HE AAC at 24kbps, over all items, it is evident from Figure 2 that the HE AAC v2 codec performs statistically significantly better than the HE AAC codec.

Comparing the MPEG-4 HE AAC v2 codec operating at 24kbps with MPEG-4 HE AAC at 32kbps, over all items, it is evident from Figure 2 that the HE AAC v2 codec performs statistically equal to the HE AAC codec at 75% of the bitrate.

6 Conclusions

The verification tests all clearly show that the Parametric Stereo enhanced HE AAC technology (High Efficiency AAC v2 profile) performs as well as MPEG-4 HE AAC Profile when the latter is operating at a 33% higher bitrate. The tests also show that for no item is the new technology worse than MPEG-4 HE AAC when both coders operate at the same bitrate.

7 References

[1] Multi stimulus test with hidden reference and anchor (MUSHRA) - EBU method for subjective listening tests of intermediate audio quality, Recommendation ITU-R BS. 1534 available at http://ecs.itu.ch

[2] Audio Subgroup, **Call for Proposals for New Tools for Audio Coding,** ISO/IEC JTC1/SC29/WG11/N3793. January 2001

[3] David Meares et al. **Report on the MPEG-2 AAC Stereo Verification Tests,** ISO/IEC JTC1/SC29/WG11/N2006. February 1998

[4] Frank Feige et al., **Revised specifications of the MPEG-2 AAC Stereo Verification Tests,** ISO/IEC JTC1/SC29/WG11/N1845. October1997

[5] Audio Subgroup, **Call for Evidence for New Tools for Audio Coding,** ISO/IEC JTC1/SC29/WG11/N3641. October 2000

[6] Audio Subgroup, **Report on the Verification Tests of MPEG-4 High Efficiency AAC,** ISO/IEC JTC1/SC29/WG11/N6009. October 2003

A.1 Test methodology

A.1.1 MUSHRA

The test in this report used the MUSHRA method described in [1]. This was developed in 1999 by EBU Project Group B/AIM, in collaboration with ITU-R Working Party 6Q. An important feature of this method is the inclusion of the hidden reference and two bandwidth limited anchor signals (7 kHz and 3.5 kHz).

A quality scale is used where the intervals are labeled "bad", "poor", "fair", good" and "excellent" as opposed to BS.1116. The value on the lower end of the scale is zero, the value on the upper end is 100. No decimals are given. This scale has the advantage to be harmonized with video quality.

The length of the sequences did not exceed 20 seconds to avoid fatiguing listeners and to reduce the total duration of the listening test.

A.1.1.1.1 Training phase

In order to get reliable results, it was mandatory to train the subjects in special training sessions in advance of the test. In preparation of the test, the subjects received both explanations and instructions about the test.

The purpose of the training phase was to allow the subject to achieve two objectives as follows:

- Become familiar with all the sound excerpts under test and their quality level ranges;
- Learn how to use the test equipment and the grading scale

During the training phase, the subject was able to listen to 4 sound excerpts (among 10 that had been selected for the tests in order to illustrate the whole range of possible qualities). The sound items to which they listen to were more or less critical depending on the bit-rate and other "conditions" used. Only test items te04, te07, te20 and te48 at all tested conditions were used for the training.

During the training phase, the subject was asked to use the available scoring equipment and evaluate the quality of the sound excerpts by inputting the appropriate scores on the continuous quality scale.

The subjects were instructed that they should not necessarily give grade "Bad" to the sound excerpt with lowest quality, or grade "Excellent" to the sound excerpt with highest quality with the exception of the hidden reference that has to be graded on top of the scale. This means, **at least on out of all test items had to be graded on top of the scale.** Beside this constraint they should use the range they find appropriate.

During the training phase the subjects were able to learn how they should interpret the audible impairments in terms of the grading scale. No grades given during the training phase were taken into account in the real tests.

The purpose of the grading phase was to input individual scores in the quality scale and to get used to the user interface. The scores should reflect the subjective judgment of the quality level for each of the sound excerpts presented. During the training phase, the subjects had to run through all the tested conditions.

The subject could discuss only the perceived artifact with the test administrator but not the specific grades in order to avoid bias in individual grading.

A.1.1.1.2 User - interface

Compared to ITU-R BS.1116, the MUSHRA method has the advantage of displaying all stimuli (conditions) for one test item. The subjects were therefore able to carry out any comparison between them directly.

The whole test was divided in two sessions, each containing only one type of conditions (mono or stereo). Figure 3 below shows the user-interface presenting one item under test. The buttons represent the reference, which is specially displayed on bottom left, and all the codecs under test, including the hidden reference and both anchor points (band-limited processed reference), called test items. Above each button, with the exception of the button for the reference, a slider was used to grade the quality of the test item according to the continuous quality scale used. For each of the items, the signals under test were randomly assigned. In addition, the test items were randomized for each subject within a session. To avoid sequential effects, each subject was running the two sessions in randomized order.

None of the subjects had the same items order and the same order in the conditions presentation.

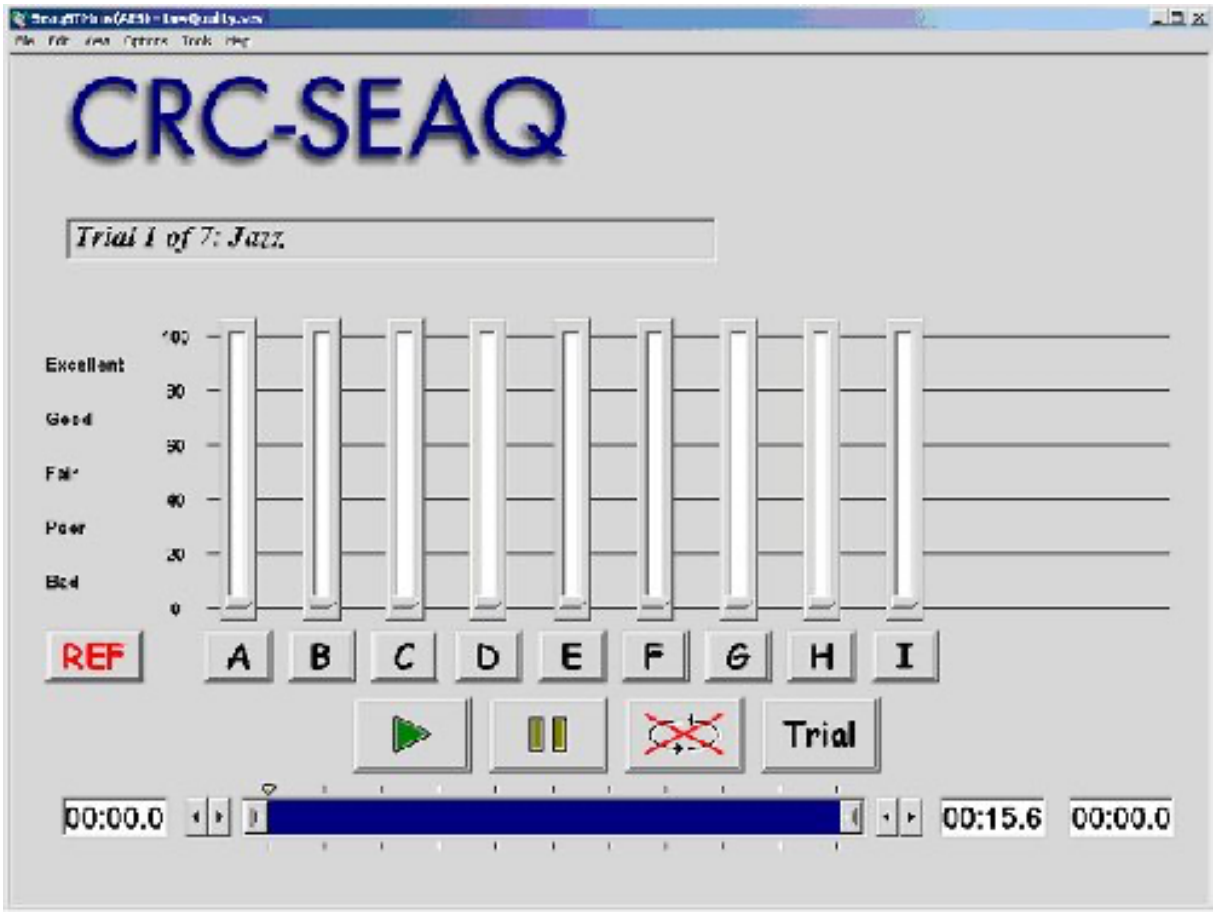

Figure 3 - User interface for MUSHRA tests

A.2 Statistical Analysis

A.2.1 General

The statistical analysis followed standard MUSHRA procedure. The calculation of the averages of the scores of all listeners remaining after post-screening will result in the Mean Subjective Scores (MSS). The first step of the analysis of the results is the calculation of the mean score $\overline{u}_{jk}$, for each of the presentations:

$$\overline{u}_{jk} = \frac{1}{N}\sum_{i=1}^{N} u_{ijk}$$

where:

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

N is the number of observers

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

$$\left[\overline{u}_{jk} - \delta_{jk}, \overline{u}_{jk} + \delta_{jk}\right]$$

where:

$$\delta_{jk} = 1.96\frac{S_{jkl}}{\sqrt{N}}$$

and the standard deviation S_{jk} is given by: $S_{jk} = \sqrt{\sum_{i=1}^{N} \frac{\left(\overline{u}_{jk} - u_{ijk}\right)^2}{\left(N-1\right)}}$.

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

Similarly, a standard deviation could be calculated for each test condition. It is noted however that this standard deviation will, in cases where a small number of test sequences are used, be influenced more by differences between the test sequences used than by variations between the assessors participating in the assessment.

A.2.2 Post-screening to assess listener reliability

For the test post screening of the listeners was done using the following post screening criterion:

Listeners should for all test-items score the three references (bandwidth limited anchor at 3.5kHz, bandwidth limited anchor at 7kHz, and the hidden reference), in ascending order. This means that the score given by the subject for the hidden reference should be larger than or equal to the score given to the 7.5kHz anchor, which should be larger than or equal to the score given to the 3.5kHz anchor.

Using these post-screening criteria, no listeners were removed from the data-set.