16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

FULLBAND CONVERSATIONAL CODEC: WHAT TESTING METHODOLOGY?

Catherine Quinquis[1], *Paolo Usai*[2]

France Telecom Technopole Anticipa, 2 Avenue Pierre Marzin 22300 Lannion, France
phone: + (33) 296051493, fax: + (33) 296053530, email: catherine.quinquis@orange-ftgroup.com
web: www.francetelecom.com
ETSI, 650, Route des Lucioles, F-06921 Sophia Antipolis, France
phone: + 33 4 92 94 42 36, fax: + 33 4 92 38 52 06, email: Paolo.Usai@etsi.org
web: http://www.3gpp.org/Support/MCC/Paolo.htm

Abstract

In ITU-T Study Group 12, the ITU-T lead study group on quality of service and performance, most of the activities have mainly focused until now on Narrowband and Wideband speech codec performance evaluation. As ITU-T Study Group 16 (that has a mandate for studies on media coding), is working on codecs with wider audio bandwidth, ITU-T SG12 has extended its scope to study quality assessment methodologies appropriate for wider bandwidths.
After a presentation of ITU-T speech and audio codec standardization process and an overview of the audio quality testing methodologies, this paper describes the ITU-T G.722.1 Fullband qualification exercise and the adopted testing methodology as well as the following characterization phase and its related methodology.

1. Introduction

ITU-T is organised in Study Groups, each of them being dedicated to some standardization areas. Study Group 16 (SG16) is responsible for studies related to multimedia service capabilities, and application capabilities; this encompasses media coding. Study Group 12 (SG12) is the lead study group on quality of service and performance in relation to the perceived quality and acceptance by users of text, data, speech, and multi-media applications. Thus, the responsibility of the standardization of speech and audio codecs, inside ITU-T, has been assigned to SG16, in collaboration with SG12.

In ITU-T SG12, most of the activities have mainly focused until now on Narrowband (300 Hz-3400 Hz) and Wideband (50 Hz-7000 Hz) speech codec performance evaluation. ITU-T SG16 began recently to work on Superwideband (50 Hz-14000 Hz) and Fullband (20 Hz-20000 Hz) audio codecs. After the successful standardization of ITU-T G.722.1 Annex C [1] [3], a low-complexity 14 kHz bandwidth audio codec extension to ITU-T Recommendation G.722.1 [2], the standardization of a Fullband extension to G.722.1 has been launched.

Subjective testing methodologies developed for Narrowband speech signals have been adapted to Wideband speech signals but it was not felt possible to extend them to wider bandwidth and other types of signals to be considered in the foreseen applications. So, some preliminary actions have been undertaken on wider bandwidth audio codec performance evaluation to assess the Superwideband and Fullband extensions to ITU-T G.722.1 Wideband codec.

The paper is organized as follows: Section 2 is dedicated to the standardization process of speech and audio codecs within ITU-T. Section 3 shortly presents the recent ITU-T G.722.1 Fullband extension standardization exercise. Section 4 gives an overview of ITU quality assessment methodologies. Sections 5 and 6 deal with audio quality evaluation performed during ITU-T G.722.1 Fullband codec qualification and characterization phases, respectively. Section 7 contains the conclusions of the paper.

2. ITU-T Codec standardization process

The ITU-T standardization process of an audio/speech codec has several stages performed in close collaboration between SG16 and SG12. The first stage is the specification of the Terms of Reference (ToR). The ToR contain the foreseen applications, the associated design constraints as well as the quality performance (expressed as requirements and objectives). Then the other stages aim to evaluate the performance of the candidate solutions to select the best codec as an ITU-T Recommendation. Three types of testing are usually successively performed: first, qualification phase test, where each candidate is individually tested to check whether it meets the requirements, then the qualified candidates are jointly tested in a selection phase test, finally a characterization phase test assesses thoroughly the quality requirements and objectives of the selected candidate.

The purpose of a qualification phase is typically to demonstrate that the performance of a candidate codec is adequate, over a minimum set of conditions representing the foreseen applications of the codec. This phase aims to pre-select, on performance basis such as quality or complexity, a subset of candidate codecs among the submitted proposals.

Then, a selection phase is run to choose one codec among the candidates, pre-selected during the qualification phase (i.e. qualified). This choice is based on performance, over a set of conditions that represent the application targeted by the codecs under consideration, on grounds of essential pre-defined requirements/objectives. This selection phase may be transformed into optimization/characterization phase if the qualified candidates agree to collaborate towards a single solution or if only one candidate passes the qualification phase.

Finally the characterization phase explores the complete scenarios of applications for the codec under test.

For each phase of the standardization exercise, the process to assess quality performance is organised in four steps. In the first step, SG16 chooses the set of ToR to be tested. The second step consists of drafting the quality assessment test plan within SG12, and the associated processing plan within SG16. During the third step, subjective and/or objective quality tests are run (c/o SG12). The fourth step, under SG12 responsibility, consists in the analysis of the quality test results and sending information/recommendation to SG16.

3. ITU-T Fullband audio codec standardization

After the successful standardization of ITU-T G.722.1 Annex C [1] [3], a low-complexity 14 kHz bandwidth audio codec, SG16 discussed the need to standardise a new codec with greater bandwidth that will be the first ITU-T Fullband audio codec. The primary applications envisaged for this codec were video- and teleconferencing with open air microphones (including speakerphones), a secondary application being Internet audio streaming. In November 2006, SG16 agreed to launch the standardization of a low-complexity Fullband audio codec (20-20000 Hz bandwidth, 48 kHz sampling frequency) at 32, 48, and 64 kbit/s for wired conversional applications. As a fast time-to-market approach was needed, it was also decided that the best way to proceed was to standardize this new low-complexity Fullband audio codec as another extension of ITU-T G.722.1.

In early 2007, the ToR were discussed and finalized and the qualification phase prepared. SG12 worked on appropriate test methodologies for fullband audio coding and designed the quality assessment test plans whereas SG16 prepared the related processing plan. Four companies declared their intention to participate in the qualification phase. Finally, only two companies submitted a candidate. In June 2007, SG12 analysed the quality test results and recommended to SG16 that both candidate codecs be allowed to go forward for the next phase. SG16 also reviewed the other performance of the two candidates (complexity, delay and frequency response) and agreed that both candidates were qualified.

While the collaboration between the two qualified candidates was under discussion, the next phase was prepared considering two options: a selection phase or an optimization/characterization phase. In mid- September 2007, the two candidates announced that they had agreed to collaborate. So an optimization/characterization phase was prepared. SG12 discussed a suitable testing methodology and, after selecting one G.722.1 Fullband extension, drafted an optimization/characterization quality assessment test plan. SG12 also provided guidelines on suitable input signals for measuring the frequency response of Fullband codecs.

4. ITU testing methodologies

The performance assessment mainly consists in evaluating the quality for the testing conditions listed in the ToR. This evaluation is generally performed with formal subjective listening tests. The procedures for conducting the tests (processing and quality assessment) have to be well defined and reflect the future usage conditions. The listening test plan design focuses on the suitable subjective test methodologies and the associated test factors (number and type of listeners, number of talkers/items, number of samples, anchors, etc…). ITU-T quality assessment tests of speech and audio coding technologies are conducted in specialized listening labs by employing panels of native listeners and by using a large number of samples/items. It must be noted that whereas naive listeners are used in the narrow and Wideband tests, experienced listeners are requested for wider bandwidth audio coding tests.

4.1. Testing Narrowband and Wideband codecs

Within ITU-T SG12, testing methodologies [4] were first developed to assess the quality of Narrowband speech codecs fitting with the Plain Old Telephone System (POTS) bandwidth. Mainly three testing methodologies were adopted for listening only test:

- Absolute Category Rating (ACR)
- Degradation Category Rating (DCR)
- Comparison Category Rating (CCR)

Such methodologies use a reference system built on grounds of the degradation observed with PCM coding (multiplicative noise), the reference system being the Modulated Noise Reference Unit (MNRU), described in ITU-T Recommendation P.810 [5].

These testing methodologies were further adapted to Wideband speech in the 90s', and the reference system was modified to accommodate Wideband signals.

4.2. Testing wider bandwidths

Adaptation of ACR and DCR testing methodologies to accommodate Fullband signals was felt rather difficult. First, foreseen applications for Fullband conversational codecs are not only speech but also mixed content or even music. Mixed content represents advertisement, ring back tones, music on hold and even film trailers. Furthermore, the reference system used by the "classic" methodologies would by all means be inappropriate for Fullband signals, and would need to be modified or even re-invented. Such a modification in a testing methodology implies at least one phase of validation. The time schedule imposed by SG16 did not allow enough time for the adaptation of the methodologies to Fullband signals and for conducting the validation phase. Therefore, it was agreed that the quality assessment method of the Fullband conversational codec would have used well consolidated ITU-R testing methodology/ies, until SG12 develop its own appropriate methodologies. ITU-R mandate includes radio-communication broadcasting, as well as vision, sound, multimedia and data services for delivery to the general public, and also technologies for quality control.

Three different testing methodologies from ITU-R were felt of interest for the qualification exercise: BS.1116 [6], BS.1285 [7] and BS.1534[8].

- The ITU-R Recommendation BS.1116 is intended for use in the assessment of systems which introduce impairments so small to be undetectable without rigorous control of the experimental conditions and appropriate statistical analysis.
- The ITU-R Recommendation BS.1285 is based on Recommendation ITU-R BS.1116, and may be used as pre-selection methodology, in case the systems under test produce significant impairments, avoiding then to carry out BS.1116 unnecessary tests.
- The ITU-R Recommendation BS.1534 is called “MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)”. The MUSHRA method is suitable for evaluation of 'intermediate' audio quality and gives accurate and reliable results.

4.3. Processing

In ITU-T audio coding standardization process, the processing of the audio material (speech and other sounds such as music, jingles, background noises, etc…) is well defined. The processing plan describes how audio material has to be prepared to simulate the real usage conditions. It specifies what and how to use software tools and shows, in the form of diagrams, the processing stages required. The set of software tools for the development of speech and audio coding standards is contained in the ITU-T Software Tool Library (STL), which is used for the processing. Since its first publication in 1992, substantial improvement and new features were introduced in successive releases. The last STL release (STL2005 [9]) incorporates tools developed for Superwideband signals processing such as a Superwideband terminal characteristics filter and a reverberation tool. Since such release in July 2005, Fullband audio signal process-

ing has required further tool adaptation or the development of new tools. For instance, a new band-pass filter simulating the acoustic input characteristics of Fullband (20-20000 Hz) terminals has been designed. An additional reverberation room impulse to be used with the reverberation tool has been provided as well. Suitable test input signals to measure frequency response of Fullband codecs have also been studied.

5. ITU-T G.722.1 Fullband Extension Qualification phase

5.1. Testing methodology for qualification

In the standardization process, the qualification phase is clearly a pre-selection phase where the candidate codec must show its promising performance. If ITU-R BS.1116 is used for systems that introduce relatively large and easily detectable impairments, it may lead to less useful results to the purpose than a simpler test. For that reason BS.1116 was discarded. BS.1534 was felt difficult to use due to its natural limitation in number of conditions to be tested. Hence BS.1285 was felt appropriate for a qualification phase. It is detailed below.

In the "triple stimulus/hidden reference/double blind" method, three audio stimuli, reference “Ref”, signal “A”, and signal “B”, are assessed by the listeners. “Ref” and one of the audio signals “A” or “B” are the reference or uncoded source material, whilst the remaining stimulus (“B” or “A”) is the coded material. The allocation of “A” and “B” to the hidden reference or the coded version is set at random. The randomisation table is provided with the test plan.

The rating scale is the five-point impairment scale below:

5.0 Impairment is imperceptible

4.0 Impairment is perceptible but not annoying

3.0 Impairment is slightly annoying

2.0 Impairment is annoying

1.0 Impairment is very annoying

The listening panel is selected among experienced listeners, who are familiar to the type of impairments that are likely to occur during the test. The subject is required to first identify which sample (A or B) is the 'Ref' sample and rate that sample a score of 5.0. The subject then rates the other sample. Ratings are entered with one decimal-point accuracy. A rejection technique is used after the real test (post-screening) in order to eliminate (eventually) non reliable subjects. Here, elimination is referred to as a process where all judgements from a particular subject are discarded and omitted from the analysis of raw data.

5.2. Quality assessment test plan design

Therefore, the quality assessment test plan for the qualification phase of the G.722.1 Fullband extension used the BS.1285 methodology. The testing was organised in three experiments:

- Experiment 1 to evaluate the codec quality under clean reverberant speech conditions
- Experiment 2 to evaluate the codec quality under noisy reverberant speech conditions, the background noise being 'interfering talker' associated to 'office noise'
- Experiment 3 to evaluate the codec quality under mixed content conditions

The ToR requested to compare the quality of the candidate codec to the LAME MP3 version 3.97 (Release date: 24 September 2006). Three bit rates were defined for the candidate codec (32, 48 and 64 kbit/s) to be compared with three bit rates of the reference codec (40, 56 and 64 kbit/s, respectively.)

5.3. Qualification test results

Each candidate in the qualification phase was tested separately against the reference codec.

Each condition 'c' was assessed for T (= 4) different talkers/sentence-pairs by L (=24) listeners, and the given ratings denoted by $X_{c,l,t}$ ('t'=1.. T, 'l'=1.. L). The Degradation Mean Opinion Score (DMOS) $Y_{c,t}$ for a talker 't' for condition 'c' was calculated for each condition was derived using the formula:

$$Y_{c,t} = \frac{1}{L}\sum_{l=1}^{L} X_{c,l,t} \,. \qquad (1)$$

The overall DMOS Y_c for all talkers for condition 'c' was then obtained from:

$$Y_c = \frac{1}{T}\sum_{t=1}^{T} Y_{c,t} \,. \qquad (2)$$

The standard deviation S_c for condition 'c' was calculated from:

$$S_c = \sqrt{\frac{1}{L \times T - 1}\sum_{t=1}^{T}\sum_{l=1}^{L}(X_{c,l,t} - Y_c)^2} \,. \qquad (3)$$

Finally, the confidence interval CI_c at the (1-α) confidence level was calculated for $N = L \times T$ from:

$$CI_c = (t_{1-\alpha,\, N-1})\frac{S_c}{\sqrt{N}} \,. \qquad (4)$$

Every requirement was tested against the reference at 95% confidence level (α=0.05).

Figures 1a, 1b, 1c show the performance of the candidates and of the reference codec LAME MP3: in clean speech (Experiment 1, Fig. 1a), in noisy reverberant speech (Experiment 2, Fig. 1b), for mixed content (Experiment 3, Fig. 1c). Codec candidates are called 'cand A' and 'cand B' in the figures.

For experiment 1, the test results show that one of the candidate codecs is better than the reference codec for the lowest bit rate and shows equivalent quality for the other two bit rates. The other candidate shows about the same quality as the reference codec for every bit rate. For experiment 2, the test results show that one of the candidate codecs is better than the reference codec for the lowest bit rate and shows the same quality for the two other bitrates. The other candidate shows about the same quality as the reference codec for every bit rate. For experiment 3, the test results show that the candidate codecs are better than the reference codec for the lowest bit rate and show the same quality for the two other bitrates.

Based on the analysis of the qualification test results and the review of the other performance, it was decided that the two candidates were qualified. As they chose to jointly work to provide only one candidate for the next phase, the next phase was an optimization/characterization phase.

6. Optimization/Characterization phase

After the analysis of the results of the qualification phase of the Fullband extension of G.722.1 the test methodology described in

ITU-R BS.1285 was felt appropriate for qualification phase. Nevertheless, a more accurate testing methodology was felt necessary for the next phases, selection and/or characterization.

Considering the performance in terms of quality of the candidate codecs observed during the qualification phase, SG12 decided to use the ITU-R BS.1116 as quality assessment methodology for the optimization/characterization phase. Two experiments were designed, one dealing with speech and the other one with music and mixed content. Two laboratories were involved in the optimization /characterization phase. Experiment 1 intended to assess the performance of the candidate with speech in the following conditions: clean speech, clean reverberant speech, and reverberant speech in presence of background noise (office noise and interfering talker). Experiment 1 was run in two different languages on a set of six talkers and 3 sentences per talker. Experiment 2 assessed the performance of the candidate with mixed content and music. Three types of mixed content items were considered: advertisement, film trailer, news with jingle. They contain speech, music, noises. Music is classified into two groups: classical and modern. Experiment 2 was run in two different languages for mixed contents, and different items for music. The test was run using three samples per type of mixed content and three meaningful musical passages per type of music. For each experiment, the listening panel was selected among experienced listeners, who are familiar to the type of impairments that are likely to occur during the test. A set of 24 listeners remained after post-screening.
The test results were statistically analysed and compared to LAME MP3 as in the previous phase.
Figures 2a, 2b show the performance of the candidates and of the reference codec LAME MP3 in clean speech, in reverberant speech and in noisy reverberant speech (Experiment 1); Figures 2c, 2d for mixed content (Experiment 2) and Figures 2e, 2f for music (Experiment 3). Codec candidate is called 'CuT' in the figures. High rate represents 64 kbit/s for the CuT and LAME MP3, mid rate is 48 kbits/s for the CuT and 56kbit/s for LAME MP3 and low rate is 32 kbit/s for the CuT and 40 kbit/s for LAME MP3.

For experiment 1, the test results show that the candidate codec is better than the reference codec for all bit rates in both languages. For experiment 2, the test results show that the candidate codec is better than the reference codec for the lowest bit rate and for the two other bitrates the quality depends on the language/items.

7. Conclusion

After a presentation of ITU-T speech and audio codec standardization process and an overview of the audio quality testing methodologies, this paper has considered a set of suitable standard methodologies that can be adopted and used for the subjective testing of Fullband conversational codecs. The standardization of the Fullband extension to G.722.1 was described and quality test results reported. As many important applications either require or will greatly benefit from audio bandwidths wider than 7 kHz, the work of testing methodology for such wider bandwidths will gain momentum in the next ITU-T study period (2009-2012).

8. Acknowlegment

The authors would like to acknowledge the Q7/12 experts, in particular the listening laboratories involved in this exercise, Dynastat and France Telecom.

References

[1] M. Xie & al, "ITU-T G.722.1 Annex C: A New Low-Complexity 14 kHz Audio Coding Standard", Proc. ICASSP, May 2006.

[2] ITU-T Rec. G.722.1, "Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss", September 1999.

[3] ITU-T Rec. G.722.1 Annex C, "Low complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss Annex C 14 kHz Mode at 24, 32, and 48 kbit/s ", May 2005.

[4] ITU-T P.800 Methods for subjective determination of transmission quality.

[5] ITU-T P.810 Modulated noise reference unit (MNRU).

[6] ITU-R BS.1116 Methods for the subjective assessment of small impairments in audio systems including multi-channel sound systems.

[7] ITU-R BS.1285 Pre-selection methods for the subjective assessment of small impairments in audio systems.

[8] ITU-R BS.1534 Method for the subjective assessment of intermediate quality level of coding systems.

[9] ITU-T G.191 STL-2005 Manual, "ITU-T Software Tool Library 2005 User's manual", 2005.

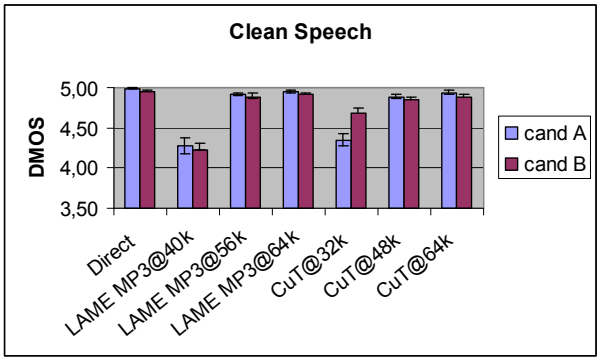

Figure 1a –ITU-T G.722.1 Fullband qualification results (Exp.1)

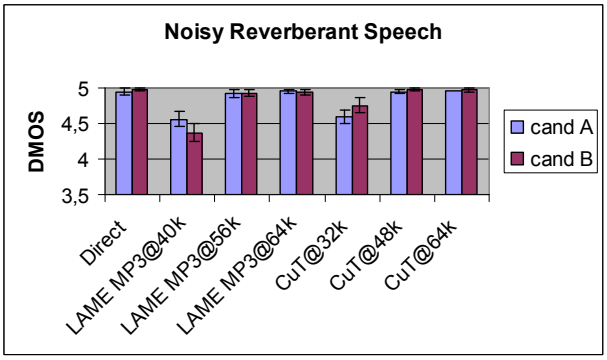

Figure 1b –ITU-T G.722.1 Fullband qualification results (Exp.2)

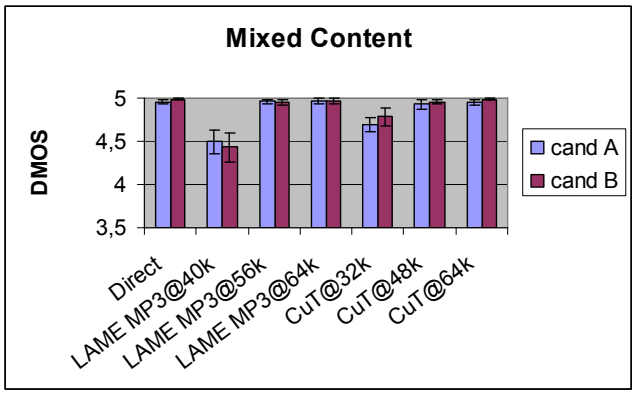

Figure 1c –ITU-T G.722.1 Fullband qualification results (Exp.3)

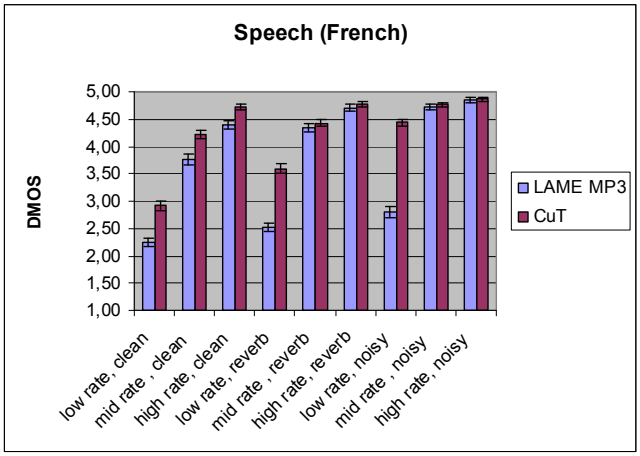

Figure 2a –ITU-T G.722.1 Fullband optimization/characterization results (Exp.1: French Language)

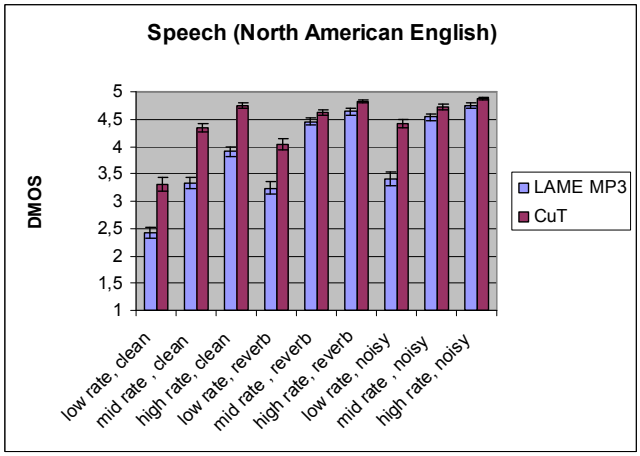

Figure 2b –ITU-T G.722.1 Fullband optimization/characterization results (Exp.1: North American Language)

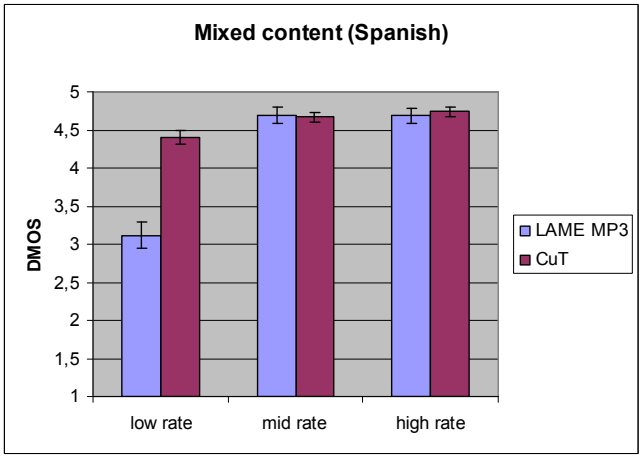

Figure 2c –ITU-T G.722.1 Fullband optimization/characterization results (Exp.2: Spanish Language)

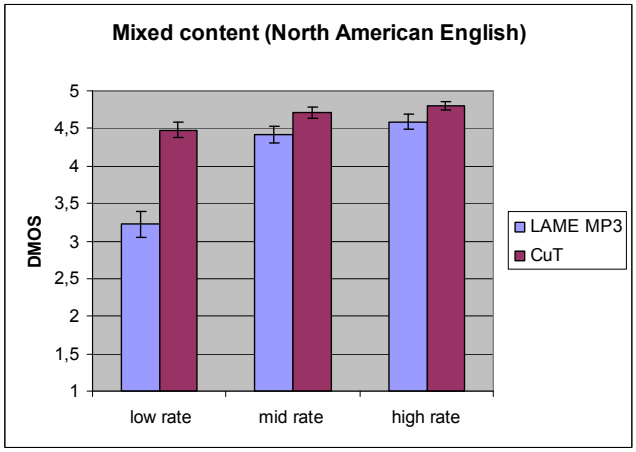

Figure 2d –ITU-T G.722.1 Fullband optimization/characterization results (Exp.2: North American Language)

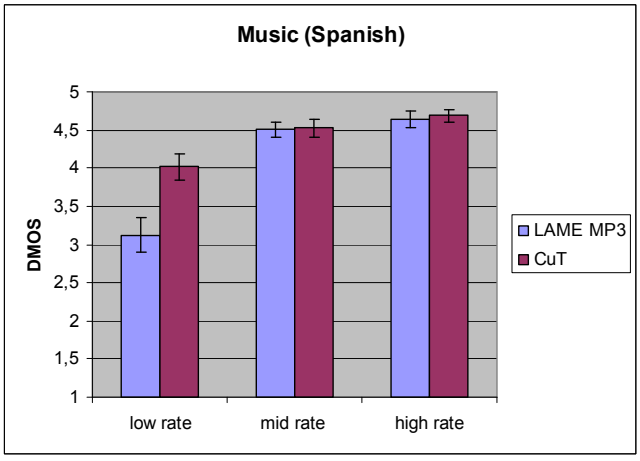

Figure 2e –ITU-T G.722.1 Fullband optimization/characterization results (Exp.2: Spanish Music)

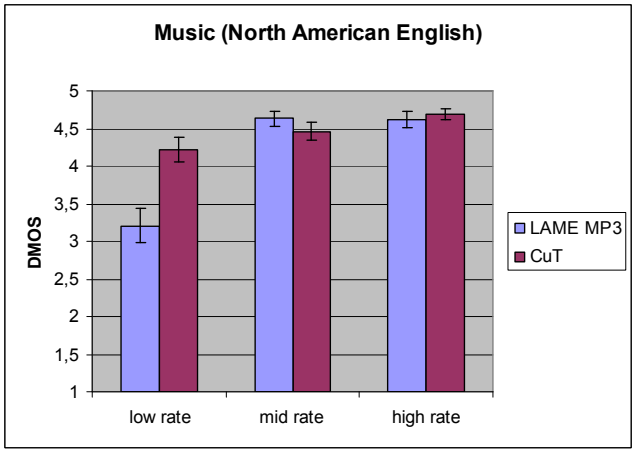

Figure 2f –ITU-T G.722.1 Fullband optimization/characterization results (Exp.2: North American Music)