ITU-T G.722.1 ANNEX C: A NEW LOW-COMPLEXITY 14 KHZ AUDIO CODING STANDARD

Minjie Xie[1], Dave Lindbergh[2], and Peter Chu[1]

[1] Polycom, Inc., 10 Mall Road, Suite 110, Burlington, MA 01803, USA
[2] Polycom, Inc., 100 Minuteman Road, Andover, MA 01810, USA
E-mail: {Minjie.Xie, Dave.Lindbergh, Peter.Chu}@polycom.com

ABSTRACT

This paper describes the low-complexity 14 kHz audio coding algorithm which has been recently standardized by ITU-T as Recommendation G.722.1 Annex C ("G.722.1C"). The algorithm is an extension to ITU-T Recommendation G.722.1 and doubles the G.722.1 algorithm to permit 14 kHz audio bandwidth using a 32 kHz audio sample rate, at 24, 32, and 48 kbit/s. The G.722.1C codec features very high audio quality and extremely low computational complexity compared to other state-of-the-art audio coding algorithms. This codec is suitable for use in video conferencing and teleconferencing, and Internet streaming applications. Subjective test results from the Characterization Phase of G.722.1C are also presented in the paper.

1. INTRODUCTION

As the video conferencing and teleconferencing markets mature, these applications are experiencing strong market and user pressure toward support of super-wideband audio quality.

In video conferencing applications, there is already a strong market trend toward 14 kHz audio as an improvement on 7 kHz wideband audio provided with G.722, G.722.1, and G.722.2. Despite the relatively small amount of energy above 7 kHz in human speech, the additional octave of bandwidth represents a very significant improvement in the subjective quality and "naturalness" of speech applications, and of course is very important for music reproduction. Low complexity is especially important in video conferencing applications, as the encoding of video is extremely compute-intensive; a simple audio algorithm can free cycles for video coding. Teleconferencing/speaker phone applications share most of the audio characteristics of video conferences. Since the price of speaker phones is very sensitive to the cost of the DSP chip, the market success of super-wideband speakerphones will greatly benefit from a low-complexity algorithm. In this market, 7 kHz wideband capability is increasingly commonly offered, and the same market pressures that led to super-wideband audio for videoconferencing are expected to drive speakerphones in the same direction very soon.

The urgency of the market need for such an audio coding standard was illustrated by products already in the market in 2004, from an increasing number of vendors, which offered super-wideband audio using non-interoperable proprietary schemes. Without a standardized low-complexity codec, full interoperability with low-cost devices can be difficult (requiring transcoding) or impossible.

At the June 2004 meeting of ITU-T Q.10/SG16, Polycom, Inc. proposed that ITU-T should standardize a low-complexity super-wideband audio codec and submitted its Siren14™ codec, a doubled form of the algorithm of G.722.1, to ITU-T as a candidate. The Siren14™ codec met all the requirements in the subjective tests for the Qualification Phase in November 2004 and for the Characterization Phase in March 2005, respectively. At the April 2005 meeting of ITU-T WP3/SG16, Siren14™ was adopted as ITU-T Recommendation G.722.1 Annex C.

In this paper, we first review the main mode of ITU-T Recommendation G.722.1 and then describe the algorithm of G.722.1 Annex C. Finally, subjective test results from the Characterization Phase of G.722.1 Annex C are summarized.

2. OVERVIEW OF MAIN G.722.1 MODE

The main mode of ITU-T Recommendation G.722.1 describes a wideband coding algorithm that provides an audio bandwidth of 50 Hz to 7 kHz, operating at a bit rate of 24 kbit/s or 32 kbit/s [1]. The coder is based on transform coding, using a Modulated Lapped Transform [2] and operates on frames of 20 ms corresponding to 320 samples at a 16 kHz sampling rate. Because the transform window length is 640 samples and a 50% overlap is used between frames, the effective look-ahead buffer size is 320 samples. Hence the total algorithmic delay of the coder is 40 ms.

2.1 Encoder

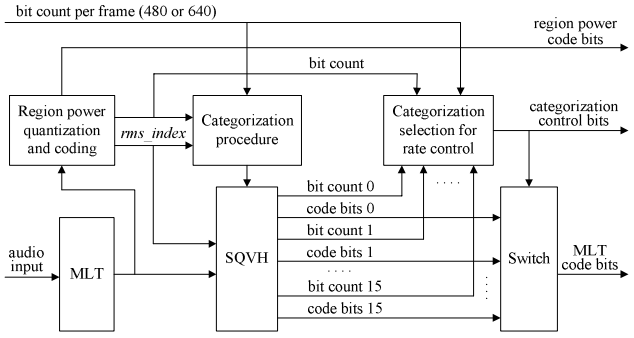

Figure 1. Block diagram of the G.722.1 encoder

Figure 1 shows a block diagram of the encoder. The Modulated Lapped Transform (MLT) performs a frequency spectrum analysis on audio samples and converts the samples from the time domain into a frequency domain representation. Every 20 ms the most recent 640 audio samples are fed to the MLT and are transformed into a frame of 320 transform coefficients centered at 25 Hz intervals. The MLT transform coefficients are divided into 16 regions, each having 20 transform coefficients, respectively. Thus, each region represents a bandwidth of 500 Hz. As the bandwidth

is 7 kHz, only the 14 lowest regions, r_0-r_{13}, are used. The 40 MLT coefficients representing frequencies above 7 kHz are ignored.

The region power or spectrum energy is determined for each region. The region power is defined as the root-mean-square (*rms*) value of the MLT coefficients in the region. The region power is then quantized with a logarithmic quantizer. The possible quantization values are the set $2^{(i/2+1)}$, where i is an integer in the range [-8, 31]. The quantization index for the lowest frequency region, *rms_index*(*0*), is further constrained to the range [1, 31]. The region power in the lowest frequency region, *rms*(*0*), is quantized with 5 bits and the quantization index *rms_index*(*0*) is directly transmitted. The quantization indices of the remaining 13 regions are differentially coded against the last highest-numbered region and then Huffman coded with a variable number of bits. The region power code bits are transmitted in order from the lowest frequency region to the highest. The region power quantization and coding module also produces the corresponding bit count for the categorization procedure described below.

Using the quantized region power indices and the number of bits remaining in the frame, the categorization procedure generates 16 possible categorizations to determine the parameters used to quantize and code the MLT coefficients. Each categorization consists of a set of 14 "category" assignments, one assignment for each of the 14 regions. The category assigned to a region defines the quantization and coding parameters such as quantization step size, dead zone, vector dimension, and variable bit-length code for that region and an expected total number of bits for representing the quantized MLT coefficients in the region. There are 8 categories: category 0 through category 7. Category 0 has the smallest quantization step size and uses the most bits. Category 7 has only one quantization output value, set to "0". Only one of the 16 possible categorizations will be selected and transmitted with 4 categorization control bits.

MLT coefficients in categories 0 through 6 are normalized, scalar quantized, combined into vectors, and Huffman coded in the Scalar Quantized Vector Huffman Coding (SQVH) module shown in Figure 1. As regions assigned a category 7 are quantized to "0", they are not allocated any bits for transmission. For each region assigned a category from 0 to 6, the MLT coefficients are first separated into sign and magnitude parts. Then the magnitude parts of the MLT coefficients are normalized by the quantized region power and scalar quantized with dead zone expansion. The resulting quantization indices are combined into vector indices and then Huffman coded. The sign bits are directly transmitted.

For each of the 16 possible categorizations, the total number of bits actually required to represent the frame is computed. This includes the bits used to represent the quantized region powers, the 4 categorization control bits, and the bits needed for the SQVH coding of the MLT coefficients. From those categorizations which yield bit totals that fit within the allotment, the categorization with the lowest index is selected. If no categorization yields a bit total that fits within the allotment, the categorization that comes closest is selected. Then, code bits are transmitted until the allotment for the frame is exhausted. The bit stream transmitted on the channel is comprised of 3 parts: region power code bits, categorization control bits, and SQVH code bits for MLT coefficients.

2.2 Decoder

Figure 2 shows a block diagram of the decoder. The data received at the decoder is demultiplexed into region power code bits, categorization control bits, and SQVH code bits for MLT coefficients. For every frame, the first 5 bits of the bit stream are decoded to obtain the quantized region power index *rms_index*(*0*). Then, for regions 1 through 13, the variable bit-length codes for differential indices are decoded and the quantization indices for these region powers are reconstructed.

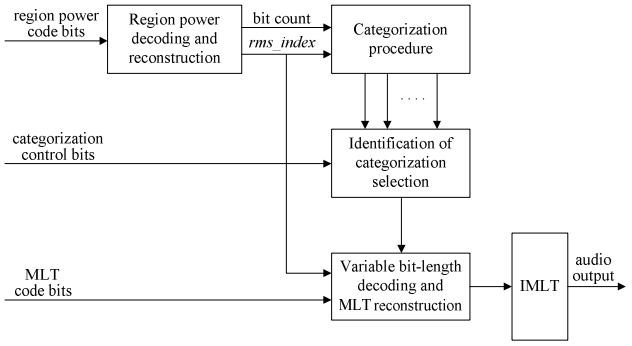

Figure 2. Block diagram of the G.722.1 decoder

After decoding the region powers, the number of bits available for representing the MLT coefficients is calculated. Using the same categorization procedure as the encoder, the set of 16 possible categorizations computed by the encoder are reconstructed. The 4 categorization control bits indicate which categorization was used to encode the MLT coefficients, and consequently should also be used by the decoder.

For each region, the variable bit-length codes representing the MLT vectors are decoded for the appropriate category. The individual MLT coefficient quantization indices in a region are recovered from the vector index. The MLT coefficient amplitudes are reconstructed by computing the product of region power in the region of interest and the centroid specified by the decoded vector index. Non-zero values have their signs set according to the sign bit. The 40 MLT coefficients representing frequencies above 7 kHz are set to "0". No MLT coefficient amplitudes are encoded for regions assigned category 7. For categories 5 and 6 the quantization step sizes are so large that most MLT coefficients are coded as "0". To avoid audible artifacts, the decoder reproduces these MLT coefficients using "noise-fill" for which these 0's are replaced with values of random sign and amplitude proportional to the region power for the region.

The reconstructed MLT coefficients are converted into time domain audio samples by an Inverse MLT (IMLT). Each IMLT operation takes in 320 MLT coefficients to produce 320 audio samples.

A frame erasure concealment procedure is incorporated into the decoder. When a frame is correctly received, the reconstructed MLT coefficients are stored in a buffer. If the decoder is informed that a frame has been lost or corrupted, it repeats the previous frame's reconstructed MLT coefficients. If the previous frame was also lost or corrupted, the decoder then sets all the MLT coefficients in the current frame to "0".

2.3 Codec Complexity

Table 1 below presents the computational complexity of G.722.1 in units of Weighted Million Operations Per Second (WMOPS). In terms of memory requirements, G.722.1 needs about 11K bytes RAM and 20K bytes ROM.

Table 1. Computational complexity of G.722.1 (WMOPS)

Bit rate (kbit/s)	Encoder	Decoder	Encoder+Decoder
24	2.3	2.7	5.0
32	2.4	2.9	5.3

3. ALGORITHM OF G.722.1 ANNEX C

ITU-T Recommendation G.722.1 Annex C [3] describes the 14 kHz mode at 24, 32, and 48 kbit/s for G.722.1. G.722.1C has the same algorithmic steps as the main G.722.1 mode, except that the algorithm is doubled to accommodate the 14 kHz audio bandwidth.

The G.722.1C codec is designed to operate with an audio signal sampled at 32 kHz. Compared to the main body of G.722.1, this codec also operates on frames of 20 ms but the frame length is doubled from 320 samples to 640 samples due to the higher sampling frequency. The transform window size increases to 1280 samples from 640 samples in G.722.1. The algorithmic delay is still 40 ms in G.722.1C.

3.1 Encoder

The specific differences in the G.722.1C encoding algorithm compared to the main G.722.1 mode are as follows:

- Double the MLT transform length from 320 to 640 samples
- Double the number of frequency regions from 14 to 28
- Double the Huffman coding tables for encoding quantized region power indices
- Double threshold for adjusting the number of available bits from 320 to 640

In G.722.1C, the MLT transforms the newest 1280 audio samples into 640 transform coefficients and is given by

$$mlt(m) = \sum_{n=0}^{639} \sqrt{\frac{2}{640}} \cos\left(\frac{\pi}{640}(n+0.5)(m+0.5)\right) v(n), \quad 0 \le m < 640 \qquad (1)$$

where

$$v(n) = w(319-n)x(319-n) + w(320+n)x(320+n), \quad 0 \le n < 320 \qquad (2)$$

$$v(n+320) = w(639-n)x(640+n) - w(n)x(1279-n), \quad 0 \le n < 320 \qquad (3)$$

and

$$w(n) = \sin\left(\frac{\pi}{1280}(n+0.5)\right), \quad 0 \le n < 640 . \qquad (4)$$

The MLT coefficients are divided into 32 regions of 20 transform coefficients. Each region spans 500 Hz. Due to the 14 kHz bandwidth, only the 28 lowest regions, or 560 MLT coefficients, are used. The 80 MLT coefficients representing frequencies above 14 kHz are ignored. For each region, the region power is computed, quantized, and coded using the same method as the G.722.1 encoder. To encode the quantization indices of 28 regions, the new Huffman coding tables are obtained by repeating the last row of the corresponding tables of G.722.1 fourteen times. The categorization procedure and SQVH coding of the MLT coefficients are performed in the same way as G.722.1. The bit stream is transmitted in the same format and order as G.722.1.

3.2 Decoder

Following are the main changes in the G.722.1C decoder compared to G.722.1:

- Double the number of frequency regions from 14 to 28
- Double threshold for adjusting the number of available bits from 320 to 640
- Extend the centroid table for reconstruction of MLT coefficients
- Double the IMLT transform length from 320 to 640 samples

As in the G.722.1 decoder, the first 5 bits of the bit stream are assembled into the quantized region power index *rms_index*(*0*). Then, the remaining 27 region powers are Huffman decoded and reconstructed. The 4 categorization control bits are then decoded to determine which of the 16 possible categorizations was selected and transmitted by the encoder. Just as in the G.722.1 decoder, the categorization procedure uses the quantized region powers together with the number of bits remaining to be decoded in the current frame and computes the set of 16 possible categorizations. The remaining code bits in the frame represent the quantized MLT coefficients and they are decoded according to the category information for each region. To reconstruct the MLT coefficients for the 28 regions, the new centroid table is obtained by adding a new row and 2 columns of "0" to the table used for G.722.1. For each region, the reconstructed MLT coefficient amplitudes are obtained by taking the product of region power with the centroid specified by the decoded vector index. The signs of non-zero amplitudes are determined by the sign bits received. The 80 MLT coefficients representing frequencies above 14 kHz are set to "0". The same technique of noise-fill that the G.722.1 decoder uses is applied to category 5 through category 7.

After reconstruction of the MLT coefficients, they are transformed into 640 time domain audio samples by an IMLT, which can be decomposed into a type IV DCT followed by a *window-overlap-add* operation. The type IV DCT is defined as

$$u(n) = \sum_{m=0}^{639} \sqrt{\frac{2}{640}} \cos\left(\frac{\pi}{640}(m+0.5)(n+0.5)\right) mlt(m) , \quad 0 \le n < 640 \qquad (5)$$

The *window-overlap-add* operation uses half of the samples from the current frame's DCT output with half of those from the previous frame's DCT output, *u_old*(*n*), as follows:

$$y(n) = w(n)u(319-n) + w(639-n)u\_old(n), \qquad (6)$$

$$y(n+320) = w(320+n)u(n) - w(319-n)u\_old(319-n), \qquad (7)$$

where $0 \le n < 320$ and $w(n)$ is defined in (4).

The unused half of the current frame's DCT output, $u(n+320)$, is stored as *u_old*(*n*) for use the next frame:

$$u\_old(n) = u(n+320), \quad 0 \le n < 320. \qquad (8)$$

3.3 Codec Complexity

The computational complexity of G.722.1C in WMOPS is given in Table 2. For memory requirements, G.722.1C needs about 18K bytes RAM and 30K bytes ROM. For purposes of comparison, Table 3 shows the computational complexity of G.722.1C versus two well-known audio codecs, 3GPP eAAC+ and 3GPP AMR-WB+ with in mono mode, at bit rates of 24 and 32kbit/s.

Table 2. Computational complexity of G.722.1C (WMOPS)

Bit rate (kbit/s)	Encoder	Decoder	Encoder+Decoder
24	4.5	5.3	9.7
32	4.8	5.5	10.3
48	5.1	5.9	10.9

Table 3. Computational complexity of G.722.1C vs. other codecs

Bit rate (kbit/s)	G.722.1C Enc. + Dec. (WMOPS)	3GPP eAAC+ Enc. + Dec. (WMOPS)	3GPP AMR-WB+ Enc. + Dec. (WMOPS)
24	9.7	40.8	80.1
32	10.3	42.6	86.7

The eAAC+ and AMR-WB+ WMOPS values were obtained by measuring the fixed-point implementations V6.0.0. Sound Quality Assessment Material (SQAM) speech material was used to benchmark the codecs. The 3GPP eAAC+ and AMR-WB+ encoders accept the audio files up-sampled to 48 kHz as input at 24 and 32kbit/s. To encode the input signal, the encoders then convert the 48 kHz sampling frequency into an internal sampling frequency, 32 kHz for eAAC+ and 25.6 kHz for AMR-WB+.

Low complexity is a major technical advantage of G.722.1C compared to other algorithms with similar performance in this bit-rate range. It represents a separate class of audio codec for low-

complexity applications such as video conferencing and teleconferencing.

4. SUBJECTIVE CHARACTERIZATION TEST RESULTS

In March 2005, as a part of the G.722.1 Annex C development process in ITU-T, subjective tests for the Characterization Phase of G.722.1C were performed by France Telecom according to a test plan [4] designed by the ITU-T SG12 Speech Quality Expert's Group (SQEG). In this section, we summarize the subjective characterization test results. More details about test design and data analysis can be found in [4]-[6].

The characterization test was comprised of two phases: the first for speech quality and the second for music and mixed content such as film trailers, news, jingles, and advertisements.

The experiments conducted in Phase 1 were divided into two main blocks:

- Experiment 1: Single talker clean speech
- Experiment 2: Single talker with background noise including interfering talkers

Experiment 1 used the Absolute Category Rating (ACR) method with the Mean Opinion Score (MOS) scale and Experiment 2 used the Degradation Category Rating (DCR) method with Degradation Mean Opinion Score (DMOS) [4]. Experiment 2 was further divided into three sub-experiments: *2a* (Reverberant speech with office noise), *2b* (Reverberant speech with interfering talkers), and *2c* (Reverberant speech with office noise and interfering talkers), to test the G.722.1C codec with different types of background noise. The MUSHRA method [4] was used in Phase 2. Experiment 3 conducted in this phase consisted of one sub-experiment for each bit rate: *3a* (24 kbit/s), *3b* (32 kbit/s), and *3c* (48 kbit/s).

All experiments were performed under error-free conditions. In each phase a floating-point version of the MPEG-4 AAC-LD codec [4] was used as the reference codec. The ITU-T requirement was that the G.722.1C codec be proven "not worse than the reference" with a 99% statistical confidence level. In addition, for purposes of comparison the G.722.1C codec was also tested against the floating-point version V6.2.0 of the 3GPP eAAC+ and AMR-WB+ codecs at the rates of 24 and 32 kbit/s.

The subjective test results of Experiments 1 and 2 in Phase 1 are presented in [5] and summarized in Table 4. From the statistical analysis of the results [5], G.722.1C met all performance requirements. In Experiments 1 and 2, G.722.1C was better than the reference codec MPEG-4 AAC-LD at 24 and 32 kbit/s and G.722.1C at 48 kbit/s was not worse than MPEG-4 AAC-LD operating at either 48 or 64 kbit/s. Compared to the 3GPP codecs, G.722.1C was not worse than 3GPP eAAC+ at 24 and 32 kbit/s and not worse than 3GPP AMR-WB+ in most of tests at 32 kbit/s.

Table 4. Subjective test results in Phase 1 (MOS/DMOS)

Condition	Exp.1	Exp.2a	Exp.2b	Exp.2c
Direct	4.44	4.55	4.74	4.59
G.722.1C at 24kbit/s	3.52	3.89	3.79	3.78
MPEG-4 AAC-LD at 24kbit/s	2.86	2.64	2.63	2.57
3GPP AMR-WB+ at 24kbit/s	4.11	4.17	4.36	4.17
3GPP eAAC+ at 24kbit/s	3.67	3.48	3.81	3.76
G.722.1C at 32kbit/s	3.80	4.18	4.24	4.16
MPEG-4 AAC-LD at 32kbit/s	3.41	3.53	3.32	3.35
3GPP AMR-WB+ at 32kbit/s	3.91	4.36	4.42	4.17
3GPP eAAC+ at 32kbit/s	3.84	3.97	4.28	4.16
G.722.1C at 48kbit/s	4.10	4.40	4.66	4.34
MPEG-4 AAC-LD at 48kbit/s	4.12	4.37	4.50	4.32
MPEG-4 AAC-LD at 64kbit/s	4.18	4.46	4.69	4.42

The subjective test results of Experiment 3 in Phase 2 are presented in [6] and summarized in Table 5. From the statistical analysis of the results [6], G.722.1C met all performance requirements. The test results show that G.722.1C was better than MPEG-4 AAC-LD at all bit rates and G.722.1C at 48 kbit/s was also better than MPEG-4 AAC-LD operating at 64 kbit/s.

Table 5. Subjective test results in Phase 2 (Mean Score)

Condition	Exp. 3a	Exp. 3b	Exp. 3c
MPEG-4 AAC-LD	28.46	46.55	59.48
G.722.1C	62.17	65.19	82.68
3GPP eAAC+	65.80	72.57	-
3GPP AMR-WB+	72.76	76.00	-
AAC-LD at 64 kbit/s	-	-	57.84
Hidden reference (original)	98.06	96.98	97.22
10kHz anchor	71.01	59.49	63.93
7kHz anchor	35.78	31.51	33.37

5. CONCLUSION

In this paper, we presented the 14 kHz audio coding algorithm and subjective characterization test results of ITU-T Recommendation G.722.1 Annex C. The G.722.1C algorithm is based on the same transform coding used in G.722.1 and is a straightforward doubling of the main mode of G.722.1. The G.722.1C codec features very high audio quality and extremely low computational complexity compared to other state-of-the-art audio coding systems. While the main application for this codec is video- and tele-conferencing with open air microphones, it can be also used for streaming audio over the Internet, and is suitable for use as a general-purpose super-wideband codec in many applications. For purposes of evaluation, the executables and audio samples are available at http://www.polycom.com/Siren14/.

6. ACKNOWLEDGMENT

The authors would like to thank Claude Lamblin, ITU-T Q.10/SG16 Rapporteur, and Catherine Quinquis, ITU-T Q.7/SG12 Rapporteur, for their great work guiding this project to a completion. The authors also thank the speech quality experts who performed the subjective characterization tests at France Telecom.

7. REFERENCES

[1] 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.

[2] H. S. Malvar, "*Signal Processing with Lapped Transforms*", Norwood, MA: Artech House, 1992.

[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 SG16 AC-05-17, "Characterization Listening Test Plan of the 14kHz Low-Complexity Audio Coding Algorithm at 24, 32 and 48 kbps Extension to ITU-T G.722.1", April 2005.

[5] ITU-T SG16 AC-05-29, "Characterization test results of the 14 kHz low-complexity audio coding algorithm at 24, 32, and 48 kbit/s extension to ITU-T G.722.1: phase 1", April 2005.

[6] ITU-T SG16 AC-05-30, "Characterization test results of the 14 kHz low-complexity audio coding algorithm at 24, 32, and 48 kbit/s extension to ITU-T G.722.1: phase 2", April 2005.