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

G.711.1: A WIDEBAND EXTENSION TO ITU-T G.711

Yusuke Hiwasaki[1]*, Shigeaki Sasaki*[1]*, Hitoshi Ohmuro*[1]*, Takeshi Mori*[1]*, Jongmo Seong*[2]*,*
Mi Suk Lee[2]*, Balázs Kövesi*[3]*, Stéphane Ragot*[3]*, Jean-Luc Garcia*[3]*, Claude Marro*[3]*,*
Lei Miao[4]*, Jianfeng Xu*[4]*, Vladimir Malenovsky*[5]*, Jimmy Lapierre*[5]*, Roch Lefebvre*[5]

[1]NTT Cyber Space Labs, Tokyo, Japan,
[2]ETRI, Daejeon, Korea,
[3]France Télécom R&D, Lannion, France,
[4]Huawei Technologies, Shenzhen, China,
[5]VoiceAge, Montreal, Canada

ABSTRACT

This paper describes a scalable coder - G.711.1 - which has been standardized by ITU-T for wideband telephony and voice over IP applications. The main feature of this extension is to give wideband scalability to ITU-T G.711, the most widely deployed speech codec. G.711.1 is designed to achieve a very short delay and low complexity. ITU-T evaluation results show that the codec fulfils all the requirements defined in the terms of reference.

1. INTRODUCTION

In March 2008, ITU-T has approved a new speech coding standard G.711.1 [1], which is an extension to ITU-T G.711 (log-compressed PCM) [2] and had been studied under the name “G.711-WB” (wideband extension). The main feature of this extension is to give G.711 wideband scalability. It aims to achieve high-quality speech services over broadband networks, particularly for IP phone and multi-point speech conferencing, while enabling a seamless interoperability with conventional terminals and systems equipped only with G.711.

This extension work-item was launched in January 2007, and the Terms of Reference (ToR) and time schedule were finalized and approved in March and June, respectively [3]. A qualification phase was first conducted to check whether candidates can pass all requirements, and five organizations participated in this phase: ETRI, France Telecom, Huawei Technologies, VoiceAge, and NTT [4][5]. This was followed by the optimization and characterization phase, where all five organizations constructively collaborated to create a unified algorithm. This paper presents the standard codec algorithm which is an outcome of this collaboration effort, and reports its quality, delay and complexity.

The paper is organized as follows: Section 2 gives a brief summary on the background of the standardization codec design, and Section 3 presents the technical details of the codec. Section 4 deals with the performance (quality, complexity, delay) of the candidate codec evaluated during the characterization phase. Finally, the paper is concluded in Section 5.

2. DESIGN CONSTRAINTS

The main emphases, put on the constraints of the coder, are as follows:

- Upward compatible with G.711 by means of embedded structure.
- The number of enhancement layers is two: a lower-band enhancement layer to reduce the G.711 quantization noise and a higher-band enhancement layer to add a wideband capability.
- Short frame-length (sub-multiples of 5 ms) to achieve low delay. The end-to-end delay over IP network must be below 150 ms [6].
- Low computational complexity and memory requirements to fit existing hardware capabilities.
- For speech signal mixing in multi-point conferences, a similar complexity to G.711 must be achieved, i.e., no increase in the complexity. It is preferable not to use inter-frame predictions, to enable enhancement layer switching in MCUs (Multipoint Control Unit) for pseudo wideband mixing, *partial mixing* [7].
- Robustness against packet losses. Preferably not too heavily dependent on interframe predictions.

With three sub-bitstreams constructed from core (Layer 0 at 64 kbit/s) and two enhancement layers (Layers 1 and 2, both at 16 kbit/s), four bitstream combinations can be constructed which correspond to four modes: R1, R2a, R2b and R3. The first two modes operate at 8 kHz sampling frequency, the last two at 16 kHz. Table 1 gives all modes and respective sub-bitstream combinations.

2.1 Partial mixing

Ordinarily, the speech mixing in conference bridges involves decoding all the coded signals from multiple locations, summation of all signals, subtraction of the signal from one’s own location, and finally re-encoding. This method is very computationally expensive especially when using wideband speech signal, but this problem can be overcome by taking advantage of a subband scalable bitstream structure, because signal can be reconstructed by decoding only part of the bitstream. In partial mixing method, only the lower-band core bitstream is decoded and mixed, and the enhancement layers are not decoded. Instead, one active location is selected among the connected locations, and its enhancement layers are redistributed to other locations. In order to implement this hybrid approach which combines redistribution and mixing, the “mixer” must judge which location to select, by means of voice-activity detection and detecting the location with the largest signal power.

This method can considerably reduce the mixing complexity required for wideband codecs, and this advantage is more significant when used on a coding scheme that operates with very low complexity: G.711. Since the core layer is continuously mixed, there will be no disruptions in the reproduced speech at the end location but only bandwidth changes. In a conferencing scenario, this is good compromise, because there is usually only one speaker at a time. However, when designing the codec, the quality degradation by the switching effect of the enhancement layers must be kept as low as possible.

Table 1: Sub-bitstream combination for each mode

Mode	Layer 0	Layer 1	Layer 2	Bit rate [kbit/s]
R1	X	-	-	64
R2a	X	X	-	80
R2b	X	-	X	80
R3	X	X	X	96

3. CODEC ALGORITHM

3.1 Overview

The codec operates on 16-kHz-sampled speech at a 5-ms frame-length. The block diagram of the encoder is shown in Figure 1. Input signal is pre-processed with a high-pass filter to remove low-frequency (0-50 Hz) components, and then split into lower-band and higher-band signals using a 32-tap analysis quadrature mirror filterbank (QMF). The lower-band signal $s_{LB}(n)$ is encoded with an embedded lower-band PCM encoder which generates G.711 compatible core bitstream (Layer 0, I_{L0}) at 64 kbit/s, and lower-band enhancement (Layer 1, I_{L1}) bitstream at 16 kbit/s. The higher-band signal $s_{HB}(n)$ is transformed into modified discrete cosine transform (MDCT) domain and the frequency domain coefficients $S_{HB}(k)$ are encoded by the higher-band encoder which generates higher-band enhancement (Layer 2, I_{L2}) bitstream at 16 kbit/s. The transform length of MDCT in the higher-band is 10 ms with a shift length of 5 ms. All bitstreams are multiplexed as a scalable bitstream.

Figure 2 shows the high-level block diagram of the decoder. The whole bitstream is de-multiplexed to G.711 compatible Layer 0, Layer 1, and Layer 2. Both, the Layer 0 and 1 bitstreams are handed to the lower-band embedded PCM decoders. The Layer 2 bitstream is given to the higher-band MDCT decoder, and decoded signal in the frequency domain $\hat{S}_{HB}(k)$ is fed to inverse MDCT (iMDCT) and the higher-band signal in time domain $\hat{s}_{HB}(n)$ is obtained. To improve the quality under frame erasures due to channel errors such as packet losses, frame erasure concealment (FERC) algorithms are applied to the lower-band and higher-band signals separately. The decoded lower- and higher-band signals, $\hat{s}_{LB}(n)$ and $\hat{s}_{HB}(n)$, are combined using a synthesis QMF filterbank to generate a wideband signal $\hat{s}_{QMF}(n)$. Noise gate processing is applied to the QMF output to reduce low-level background noise. At the decoder output, 16-kHz-sampled speech, $\hat{s}_{WB}(n)$, or 8-kHz-sampled speech, $\hat{s}_{NB}(n)$, is reproduced.

The codec has a very simple structure to achieve high quality speech with a low complexity, and is deliberately designed without any inter-frame prediction, to increase the robustness against frame erasures and to avoid annoying artefacts when enhancement layers are switched, which is required for the *partial mixing* in wideband MCU operations.

3.2 Lower-band embedded PCM codec

Figure 3 gives the block diagram of the embedded PCM encoder. It is made of the lower-band core encoder Q_{L0}, decoder Q^{-1}_{L0}, enhancement layer encoder $Q_{L1,}$ the calculator of perceptual filter coefficients a_j and filtering $F(z)$.

The lower-band core codec is based on the ITU-T G.711 standard and both μ-law and A-law companding schemes are supported. In order to achieve the best quality, the quantization noise of Layer 0 (G.711-compatible core) is shaped with a perceptual filter [8] and added to the input signal $s_{LB}(n)$ prior to quantization. This noise feedback loop is intended to improve the quality of the core PCM quantizer. This noise feedback loop is further attenuated for extreme signal conditions such as low-level input or energy concentrated at frequency close to 4 kHz. The noise shaping filter is derived from the reconstructed Layer 0 signal $\hat{s}_{L0}(n)$ by means of linear prediction (LP) analysis of order L=4:

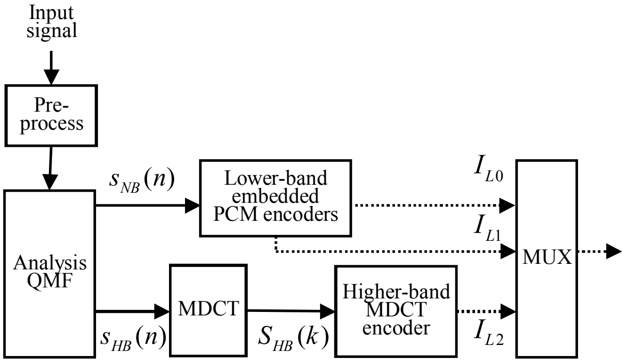

Figure 1: High-level encoder block diagram

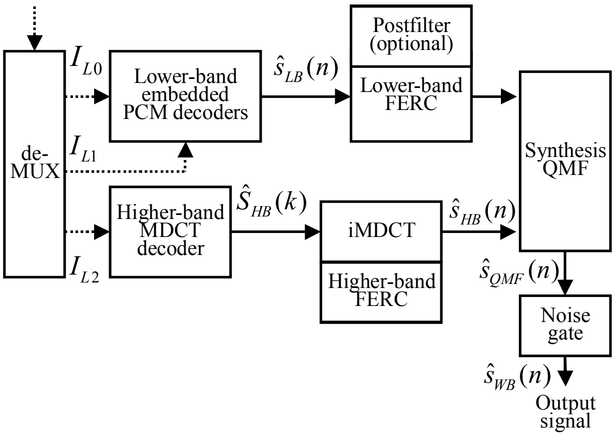

Figure 2: High-level decoder block diagram

$$F(z) = A(z/\gamma) - 1 = \sum_{j=1}^{L} \gamma^j a_j z^{-j} \ , \tag{1}$$

where $A(z)$ is the LP filter, γ is a weighting factor. A typical value of γ is 0.92. The LP filter is calculated once in every 5-ms frame and is identical at the encoder and at the decoder since it is calculated using $\hat{s}_{L0}(n)$. In the decoder, the filter $F(z)$ is only applied to the decoded signal of Layer 1 and added to the decoded signal of Layer 0. In this way the noise is properly shaped in the synthesized signal when both, Layer 0 and Layer 1 are used simultaneously.

For very-low level input signal, the weighting factor γ is attenuated to weaken the noise feedback and avoid saturation problems. To further increase the quality of synthesized signal at low level, a dead-zone quantizer is applied instead of the embedded lower-band encoder and decoder. The “dead-zone” refers to an input signal in the range [-7:7] for μ-law and [-11:11] for A-law. The purpose of the dead-zone is to enlarge the zero-output zone in the quantization. In this way, the amount of granular noise in the synthesized signal is decreased for low level signals. Further, at the decoder, an algorithm called “noise gate” is used after signal synthesis for low-level signals. This noise gate attenuates segments with power below certain threshold and as a result, the amount of low-level background noise is reduced. This improves further the perceived quality of the output signal in low-level conditions.

In order to provide a finer resolution to the core layer, the lower-band enhancement layer (Layer 1) Q_{L1} encodes the refinement signal $s'_{LBref}(n)$ using adaptive bit-allocation based on its exponent value $s'_{LBexp}(n)$. The refinement and the exponent value of each

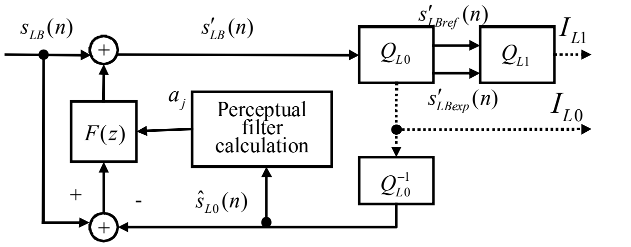

Figure 3: Lower-band embedded PCM encoder block diagram

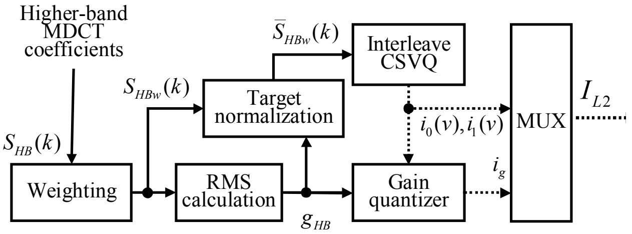

Figure 4: Higher-band MDCT encoder block diagram

sample are basically calculated by:

$$\begin{aligned} e &= \lfloor \log_2(s'_{LB}(n)) \rfloor \\ s'_{LBref}(n) &= \lfloor 2^{-e} s'_{LB}(n) \rfloor \otimes 0\text{x}07\,, \\ s'_{LBexp}(n) &= e \end{aligned} \qquad (2)$$

where $\lfloor x \rfloor$ and $\otimes$ denote rounding of x towards minus infinity and logical AND bit-operator, respectively. The constant with “0x” prefix means that the value is notated in hexadecimal.

The refinement signal has a 3-bit resolution per sample whereas the bit budget of the Layer 1 is 16 kbit/s, i.e., 80 bits per frame or 2 bits per sample. For bit reduction, the adaptive multiplexing dynamically allocates bits to each sample depending on its exponent value $s'_{LBexp}(n)$. This is possible because the exponent values are available in both encoder and decoder. The encoding is done in two stages: bit allocation table generation and refinement signal multiplexing. First, the exponent values in a frame are expanded to an exponent map, $M_{exp}(j,n)$, which stores the indices of refinement signal that uses a specific exponent index j, ($j = 0,\ldots,9$) to express the refinement signal. Simultaneously, a number of samples with the same exponent index, $N_{exp}(j)$, is counted.

1. Initialize $N_{exp}(j) = 0$ for $j = 0,\ldots,9$
2. Iterate the following Step 3 to 5 for all n samples in a frame
3. Calculate exponent index by: $j = s'_{LBexp}(n) + i,\ i = 0,1,2.$
4. Update exponent map as $M_{exp}(j, N_{exp}(j)) = n,$
5. Increment the number of samples that fit in the exponent index $N_{exp}(j) = N_{exp}(j) + 1$.

Then, the bit allocation table $B_A(n)$, i.e. the number of bits allocated to n-th sample, is computed as:

1. Initialize $B_A(n) = 0$ for $n = 0,\ldots,39$, $b^{[0]} = 80, j = 9, i = 0$
2. Calculate the number of available bits as $q = \min\lfloor b^{[i]}, N_{exp}(j)\rfloor$
3. $B_A(n) = B_A(n) + 1$ for $n = M_{exp}(j,k)$, $k = 0,..,q-1$
4. Update remaining bit-budget as $b^{[i+1]} = b^{[i]} - q$
5. if $b^{[i+1]} = 0$, done, else $j = j - 1$, $i = i + 1$, then go to Step 2

Table 2: Bit-allocation of Layer 2

Parameter	Bits per subvector	Bits per frame
MDCT coefficients (VQ)	5+5	60
Polarity (Sign)	1+1	12
Gain	-	8
Total	12	80
Bit-rate	16.0 kbit/s	

Finally, the refinement codes are calculated from the most significant bits of the refinement signal, and then are sequentially multiplexed in the Layer 1 bitstream. In the decoder, the bit allocation table is reconstructed by the same procedure as described above, and the refinement signal is reconstructed from the Layer 1 bitstream using the bit allocation table.

3.3 Higher-band MDCT codec

The higher-band MDCT coefficients are quantized using *interleaved Conjugate-Structured VQ* (CS-VQ) [5]. The details of the higher-band encoder are shown in Figure 4. Firstly, the MDCT coefficients $S_{HB}(k)$ are weighted with a set of fixed coefficients, and then normalized using the root mean square (RMS). In the interleaved CS-VQ, the weighted and normalized MDCT coefficients $\bar{S}_{HBw}(k)$ are decimated into 6 sets of 6-sample sub-vectors and those vectors are then independently quantized as 6 sub-vectors $\mathbf{S}'_{HB}(v)$ ($v = 1,\ldots,6$). This method has an advantage that adaptive bit-allocation is not required, because same number of bits can be assigned to each sub-vector. To reduce the codebook memory space, a set of conjugate-structured two-channel codebooks $\mathbf{C}_{H0w}$ and $\mathbf{C}_{H1w}$ is used, in which the decoded vector is calculated as an average of two code-vectors. A pre-selection is performed to select candidates which minimize the Euclidian distance between target sub-vector and code-vector to reduce complexity. In the pre-selection, 8 candidates are selected among 32 code-vectors in each codebook channel. After pre-selection, the best pair-indices are selected among all combination pairs of pre-selected vectors to minimize the following distance:

$$d_{HB}(v) = \left\| \mathbf{S}'_{HB}(v) - \frac{\mathbf{C}_{H0w}(i_0(v)) + \mathbf{C}_{H1w}(i_1(v))}{2} \right\|^2, \qquad (3)$$

where $\mathbf{C}_{H0w}(i_0(v))$ and $\mathbf{C}_{H1w}(i_1(v))$ are the code-vectors selected from the first and the second codebook channels, respectively for v-th normalized sub-vector $\mathbf{S}'_{HB}(v)$. By disregarding the constant terms, the above equation can be re-written as

$$d'_{HB}(v) = -\frac{2\left(\mathbf{C}_{H0w}{}^{t}\mathbf{C}_{H1w}\right) + \left\|\mathbf{C}_{H0w}\right\|^2 + \left\|\mathbf{C}_{H1w}\right\|^2}{4} + \mathbf{S}'_{HB}{}^{t}\left(\mathbf{C}_{H0w} + \mathbf{C}_{H1w}\right), \qquad (4)$$

where t denotes transposition of a vector, and the sub-vector index v and the codebook indices $i_0(v)$ and $i_1(v)$ are left out from the equation. Here, complexity is reduced by calculating the power of code-vectors, i.e., $\|\mathbf{C}_{H0w}\|^2$ and $\|\mathbf{C}_{H1w}\|^2$ and their inner products $\mathbf{C}^{t}_{H0w}\mathbf{C}_{H1w}$ beforehand and looking-up as table entries.

The frame gain g_{HB} is calculated as:

$$g_{HB} = \frac{\sum_{v=0}^{5} \mathbf{S}'_{HB}{}^{t}(v)\left(\mathbf{C}_{H0w}(i_0(v)) + \mathbf{C}_{H1w}(i_1(v))\right)}{\sum_{v=0}^{5} \left\|\mathbf{C}_{H0w}(i_0(v)) + \mathbf{C}_{H1w}(i_1(v))\right\|^2} g_{rms}\,, \qquad (5)$$

where g_{rms} is the RMS used for the normalization of the input MDCT coefficients. Then g_{HB} is compressed using μ-law and is uniformly scalar quantized with 8 bits into i_g. All indices are multiplexed to generate Layer 2 bitstream I_{L2}. Table 2 shows the bit-allocation of the higher-band encoder.

In the decoder, decoded sub-vectors are calculated as an average of two code-vectors, multiplied by the decoded gain:

$$\hat{\mathbf{S}}'_{HB}(v) = \hat{g}_{HB} \frac{\mathbf{C}_{H0w}(i_0(v)) + \mathbf{C}_{H1w}(i_1(v))}{2}, \qquad (6)$$

where $\hat{\mathbf{S}}'_{HB}(v)$ is the v-th sub-vector and $\hat{g}_{HB}$ is the decoded frame gain. All $\hat{\mathbf{S}}'_{HB}(v)$ are then interleaved to reconstruct a full set of MDCT coefficients and transformed back into time-domain by inverse-MDCT to generate higher-band signal output $\hat{s}_{HB}(n)$.

3.4 Lower-band FERC

To conceal frame erasures in the lower band, an improved version of the lower-band FERC algorithm of G.722 App. IV [9] is used. When a frame is erased, the decoder performs two steps:

- The past lower band signal is analyzed to estimate parameters including linear-predictive coding (LPC) coefficients, pitch and signal class (voiced, weakly voiced, unvoiced, transient).
- The missing signal frame is synthesized using LPC-based pitch repetition and adaptive muting. Once a good frame is received, the extrapolated signal that replaces the last erased frame is re-synchronized and cross-faded with the decoded signal. Before cross-fading, dynamic energy scaling is performed on the extrapolated signal.

3.5 Higher-band FERC

The higher-band FERC differentiates between pitch-like higher-band signal and noise-like higher-band signal in order to minimize potential quality impairments in the recovered signal generated for concealment.

When previously decoded higher-band signal exhibits a high correlation, the higher-band pitch lag is estimated around the pitch parameter calculated for the lower-band FERC. The samples of the previous pitch period in the higher-band FERC history buffer are used as the iMDCT signal of current erased frame. Then, a sine window is applied. Otherwise, an attenuated iMDCT signal from the last good frame is used. Finally, overlap-and-add is performed to generate the reconstructed higher-band signal.

3.6 Optional postfilter (Appendix I of G.711.1)

An optional postfilter designed to reduce the lower-band quantization noise at the decoder has been standardized as Appendix I of G.711.1. It enhances the quality of a 64-kbit/s bitstream when communicating with a legacy G.711 encoder.

The underlying algorithm uses *a priori* information on the properties of legacy G.711 quantization to estimate the quantization noise and derive a time-varying filter enhancing the decoded signal. For each sample coded with G.711, the quantization noise is assumed to be additive with a variance depending on the input signal energy. Since the quantization noise variance is usually very small (except for the low-level input samples), the algorithm assumes that the energy of the PCM-decoded signal is a fair estimate of the energy of the lower-band input signal.

The postfilter is estimated in frequency domain using a short-term Fourier transform of 64 samples. The G.711 quantization noise power spectral density (PSD) is estimated from the energy of the lower-band decoded signal. Then a 33-tap Wiener filter is derived using a "two-step" procedure [10]. This filter assumes that the G.711 quantization noise is white. Note that the filter estimation is robust against limited deviations from the white noise assumption. Moreover, the robustness is increased by some *a posteriori* logic which avoids excessive attenuation and limits the distortion due to

Table 3 Overview of the characterization test

Signals	Exp.	Meth.	BW	Languages
Clean speech	1a	ACR	NB	Korean, North-American English
	1b	ACR	WB	French, Chinese
Music	2a	ACR	NB	Japanese, Chinese (music)
	2b	ACR	WB	Japanese, Chinese (music)
Noisy speech	3	DCR	NB	Japanese, Korean
	4	DCR	WB	French, North-American English
Mixed speech	5a	ACR	NB	Korean, North-American English
	5b	ACR	WB	French, Chinese

Table 4: Characterization test results

CuT mode *	Reference	Exp	Condition	Score $_{CuT}$ lab A	Score $_{Ref}$ lab A	Score $_{CuT}$ lab B	Score $_{Ref}$ lab B	R/O
R1	G.711 A-law	Exp1a	Clean Speech	4.41	3.16	4.05	2.91	Req.
			3% Random FER	4.26	3.07	3.92	2.80	Req.
		Exp2a	Music	3.86	3.77	3.47	3.30	Req.
		Exp3	Background music	4.77	4.56	4.58	4.35	Req.
			Office noise	4.82	4.77	4.68	4.68	Req.
			Babble noise	4.74	4.68	4.61	4.48	Req.
			Interfering talker	4.64	4.52	4.62	4.43	Req.
R2a	16bit PCM	Exp1a	Clean Speech	4.45	4.11	4.40	4.38	Obj.
	G.711 A-law		3% Random FER	4.35	3.07	4.23	2.80	Req.
	16bit PCM	Exp2a	Music	3.85	3.90	3.46	3.43	Obj.
		Exp3	Background music	4.80	4.82	4.80	4.81	Obj.
			Office noise	4.83	4.83	4.73	4.73	Obj.
			Babble noise	4.76	4.78	4.80	4.80	Obj.
			Interfering talker	4.72	4.73	4.82	4.85	Obj.
	G.726	Exp5a	Mixed speech	4.45	2.96	4.12	2.44	Req.
R2b	G.722 56k	Exp1b	Clean Speech	4.10	3.70	4.03	3.36	Req.
			3% Random FER **	4.08	3.17	3.88	2.52	Req.
		Exp2b	Music	4.06	3.45	3.66	2.99	Req.
		Exp4	Background music	4.53	4.25	4.38	3.73	Req.
			Office noise	4.64	4.46	4.48	3.85	Req.
			Babble noise	4.71	4.41	4.56	3.79	Req.
			Interfering talker	4.61	4.48	4.68	3.89	Req.
	G.722 48k	Exp5b	Mixed speech	4.09	3.09	4.02	2.66	Obj.
R3	G.722 64k	Exp1b	Clean Speech	4.41	3.73	4.23	3.31	Req.
			3% Random FER **	4.31	3.20	4.06	2.59	Req.
		Exp2b	Music	3.91	3.56	3.75	3.07	Req.
		Exp4	Background music	4.71	4.42	4.61	3.77	Req.
			Office noise	4.68	4.51	4.51	3.98	Req.
			Babble noise	4.79	4.52	4.62	3.82	Req.
			Interfering talker	4.78	4.56	4.69	4.01	Req.
	G.722 48k	Exp5b	Mixed speech	4.28	3.09	4.23	2.66	Req.

*CuT core in each condition was G.711 A-law, ** Random FER for the reference G.722 was set to 1%.

estimation errors. The Wiener filter is applied in time domain using overlap-save method combined with filter interpolation.

4. PERFORMANCE EVALUATION

4.1 Subjective evaluation

In order to evaluate the subjective quality of the speech reproduced by the algorithm, a set of formal subjective listening tests called "Characterization test" was conducted, according to the processing and the quality assessment test plans designed and approved by ITU-T Q7/12 [11]. Several experiments were run: each twice in two different languages using 32 naive listeners, all native speakers of the respective languages. Four kind of input signals were considered: clean speech, music, noisy speech with 4 types of background noise at various SNRs (background music at 25 dB SNR, office noise at 20 dB SNR, babble noise at 30 dB; interfering talker at 15 dB SNR), and mixed speech. Both μ-law and A-law were tested.

Table 3 gives an overview of the characterization test indicating for each experiment (Exp.): the test methodology (Meth.) used where ACR (resp. DCR) denotes absolute (res. degradation) category rating, the audio bandwidth (either narrow band (NB) or wideband (WB)) and the languages used. It should be noted that for testing mixed speech conditions, the partial mixing was used for G.711.1 and was tested against full conventional mixing of the reference coder.

Table 4 gives a subset of the mean opinion score (MOS) of the tested conditions limited to -26 dBov input signal (the complete set of results is given in [12]). In this table, "CuT mode" means the test mode of the "coder under test" (i.e., the G.711.1 coder), "Reference" means the reference condition of the requirement/objective, "Score$_{\text{CuT}}$" and "Score$_{\text{Ref}}$" are the MOS of the coder and the reference coders respectively, R/O indicates whether it is a Requirement or an Objective. The judgments were made based on the statistical comparison between MOS of the codec candidate and the reference codecs, by means of a simple paired t-test at 5% significance level. The codec met all requirements and all objectives, except in objective condition of R3 high-level input (-16 dBov) in French language.

4.2 Complexity and delay

Table 5 gives the complexity and required memory of the codec for speech samples used in the above subjective evaluation. The complexity of the codec, which is estimated using basic operator set in the ITU-T Software Tool Library v2.2, is 8.70 WMOPS (Weighted Million Operations Per Second) in the worst case. This meets the ToR objective ("less than 10 WMOPS"), and when compared with another wideband extension of a narrowband codec, G.729.1 [13] (35.8 WMOPS), this figure is considerably low. The memory size of the candidate codec is 3.04 kWords RAM and 2.21 kWords table ROM, and both figures also met the memory requirements in the ToR.

The total of analysis and synthesis delays of the split-band QMF is 1.875 ms, and the delay due to the MDCT analysis for the Layer 2 is 5 ms. The overall algorithmic delay adds up to 11.875 ms (190 samples at 16 kHz), including the frame length (5 ms).

Table 5: Complexity and memory estimation

		Enc	Dec	Total
Complexity (WMOPS)	no FER	5.40	2.33	7.73
	3% FER		3.30	8.70
Memory (kWords)	Static RAM	0.18	1.50	1.68
	Scratch RAM	0.66	0.70	1.36
	Table ROM	2.21		
	Program ROM	1.94		

5. CONCLUSION

The algorithm of ITU-T G.711.1, a wideband scalable codec of G.711 proposed by ETRI, France Telecom, Huawei Technologies, VoiceAge and NTT, was described. The bitstream has an embedded structure where the core layer is generated by a G.711-compatible codec utilized with a noise shaping feedback. On top of the core layer, there are two enhancement layers: a lower band enhancement layer for the refinement signal encoded with a dynamic bit-allocation, and another one for higher band encoded with an interleaved CSVQ in MDCT domain. The emphasis in the codec design was on complexity. Formal subjective tests showed that the subjective quality of the codec met all requirements specified in the ToR in five languages. Complexity evaluation proved that a computational complexity and memory size also met the ToR objective and requirement, respectively.

ACKNOWLEDGMENTS

The authors wish to thank Alexandre Guérin for help in fixed-point conversion of the R1 postfilter and Hyun Woo Kim for helping the lower-band enhancement layer.

REFERENCES

[1] ITU-T, Geneva, Switzerland, ITU-T G.711.1 - Wideband embedded extension for G.711 pulse code modulation (pre-published), Mar. 2008.

[2] ITU-T, Geneva, Switzerland, ITU-T G.711 - Pulse code modulation (PCM) of voice frequencies, Nov. 1988.

[3] ITU-T SG16 TD 283/WP3 Annex Q10.H, "Terms of Reference (ToR) and Time schedule for ITU-T wideband extension to G.711", Study Period 2005-2008, Geneva, June 2007 (Source Q.10/16 Rapporteur).

[4] B. Kovesi, S. Ragot, and A. Le Guyader, "A 64-80-96 kbit/s Scalable Wideband Speech Coding Candidate for ITU-T G.711-WB Standardization," in *Proc. ICASSP*, pp.4801-4804, Las Vegas, U.S.A., Apr. 2008.

[5] Y. Hiwasaki, T. Mori, S. Sasaki, H. Ohmuro, and A. Kataoka, "A Wideband Speech and Audio Coding Candidate for ITU-T G.711WBE Standardization," in *Proc. ICASSP*, pp.4017-4020, Las Vegas, U.S.A., Apr. 2008.

[6] ITU-T, Geneva, Switzerland, ITU-T G.114 App. II - Guidence on one-way delay for Voice over IP, Sep. 2003

[7] Y. Hiwasaki, H. Ohmuro, T. Mori, S. Kurihara, and A. Kataoka, "A G.711 Embedded Wideband Speech Coding for VoIP Conferences," *IEICE Trans. Inf. & Syst.*, vol. E89-D, no.9, pp.2542¯2551, Sep. 2006.

[8] J. Lapierre, R. Lefebvre, B. Bessette, V. Malenovsky, R. Salami, "Noise shaping in an ITU-T G.711-Interoperable embedded codec," to appear in *Proc. 16th EUSIPCO*, Lausanne, Aug., 2008.

[9] B. Kovesi, S. Ragot, "A low complexity packet loss concealment algorithm for ITU-T G.722," in *Proc. ICASSP*, pp.4769-4772, Las Vegas, U.S.A., Apr., 2008.

[10] C. Plapous, C. Marro, P. Scalart, L. Mauuary, "A two-step noise reduction technique," *Proc. ICASSP*, vol. 1, pp.289-292, Montreal, Canada, May 2004.

[11] ITU-T SG12 TD 80/WP1, "G.711 WB extension Optimisation/ Characterization Quality Assessment Test Plan", Study Period 2005-2008, Geneva, Oct. 2007 (Source Q.7/12 Rapporteurs).

[12] ITU-T SG16 TD 479/Gen, "LS on speech and audio coding matters", Study Period 2005-2008, Geneva, Oct. 2007 (Source Q.7/12 Rapporteurs).

[13] ITU-T, Geneva, Switzerland, ITU-T G.729.1 - G.729 based Embedded Variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729, May 2006.