A WIDEBAND SPEECH AND AUDIO CODING CANDIDATE FOR ITU-T G.711WBE STANDARDIZATION

Yusuke Hiwasaki, Takeshi Mori, Shigeaki Sasaki, Hitoshi Ohmuro, and Akitoshi Kataoka

NTT Cyber Space Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo 180-8585, Japan

ABSTRACT

This paper describes an ITU-T G.711 embedded wideband speech coder, submitted as a candidate to the ITU-T G.711Wideband Extension standardization qualification phase. The codec generates a bitstream comprised of three layers: a G.711 core layer with noise shaping, a time-domain weighted vector quantized narrowband enhancement layer, and an MDCT-based higher band enhancement layer. Through subjective evaluations, the coder was found to meet all tested requirements and objectives set in terms of reference, with a low computational complexity at 10 WMOPS.

Index Terms— Wideband Speech Coding, ITU-T G.711, Pulse Code Modulation,

1. INTRODUCTION

ITU-T SG 16 is now studying the development of an extension to ITU-T G.711 (log-compressed PCM) [1], called "G.711WBE" (Wideband Extension). The main feature of this extension is to give G.711 with a wideband scalability. Its goal is 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 upon a proposal by NTT, motivated by the strong market needs in Japan. The Terms of Reference (ToR) and time schedule were finalized and approved in March and June, respectively [2]. A qualification phase was first conducted to check whether candidate can pass all requirements, and five organizations participated in this phase: ETRI, France Telecom, Huawei Technologies, VoiceAge Corp, and NTT. This paper presents a candidate codec algorithm submitted by NTT, and reports its subjective and complexity performances.

This paper is organized as follows: Section 2 gives a brief summary on motivations and codec design of the extension, and Section 3 presents the technical details of the proposed candidate codec. Section 4 deals with the subjective evaluation results of the candidate codec performed for the qualification phase, and Section 5 provides the complexity evaluation of this codec. Finally, the paper is concluded in Section 6.

2. CODEC DESIGNS

With a rapid growth in use of IP-based broadband networks, the legacy Public Switched Telephone Networks (PSTN) is being merged to and replaced by such networks. While numerous wideband speech coding algorithms have successfully been standardized, its use over broadband network has been limited. One of the reasons is that the majority of legacy PSTNs are still in use and so is G.711, and those new wideband codecs usually require costly transcoding in interoperating with legacy networks. Until the wideband speech terminals totally replace the narrowband ones, the two types of terminals will continue to co-exist. Based on the above, the main emphasis on the constraints of the coder is as follows:

- Upper compatible with G.711 by means of embedded structure.
- The number of enhancement layers is two. A lower-band enhancement layer reduces the quantization noise of the G.711 and a higher-band enhancement layer adds a wideband capability. The bitrate for those layers is set to 16 kbit/s.
- Short frame-length (sub-multiples of 5 ms) to achieve low delay. The end-to-end delay over IP network must be kept less than 150 ms.
- Low computational complexity and memory requirements to fit existing hardware capabilities.
- For speech signal mixing in multi-point conferences, a comparable 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 for pseudo wideband mixing, *partial mixing* [3].
- Robust against the packet losses. Preferable not too heavily dependant on interframe predictions.

3. CODEC ALGORITHM

3.1. Overview of the proposed codec

The candidate codec is based on a proprietary codec called UEMCLIP (mU-law EMbedded Codec structure for Low delay IP voice communication) [3]. It operates on 16-kHz-sampled speech at a 5-ms frame-length. The block diagram of the encoder is shown in Fig. 1. Input signal is preprocessed with a high-pass filter to remove low-frequency (0-50 Hz) component, and then is split into lower-band and higher-band signals using an analysis quadrature mirror filterbank (QMF). The lower-band signal is encoded with the core encoder and then the quantization residual is encoded with the lower-band enhancement sub-encoder to improve the quality in the lower band. The lower-band enhancement is based on time-domain weighted vector quantizer (VQ). The higher-band signal is encoded by a sub-encoder utilized with a frequency-domain VQ. All bitstreams, the core and the enhancements, are multiplexed as a scalable bitstream.

The block diagram of the decoder is shown in Fig. 2. Here, each sub-bitstreams are decoded by respective sub-decoders. To improve the quality in frame erasure (FER) conditions, packet-loss concealment (PLC) algorithms are applied to the lower-band and

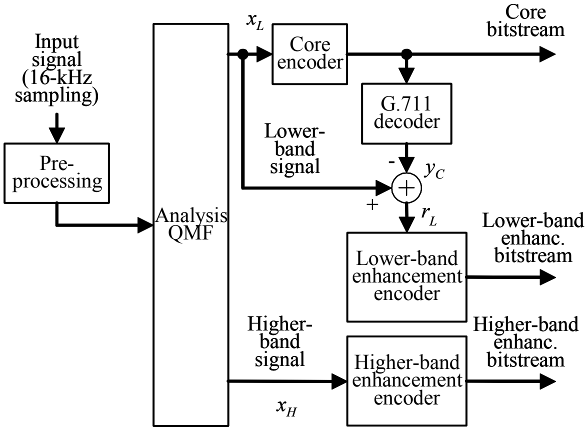
Input signal (16-kHz sampling)
Pre-processing
Analysis QMF
x_L
Core encoder
Core bitstream
G.711 decoder
Lower-band signal
-
y_C
+
r_L
Lower-band enhancement encoder
Lower-band enhanc. bitstream
Higher-band signal
x_H
Higher-band enhancement encoder
Higher-band enhanc. bitstream

Fig. 1. Encoder block diagram.

higher-band signals independently. The lower-band and higher-band signals are combined using a synthesis QMF to produce a 16-kHz-sampled output.

With three sub-bitstreams, four patterns of bitstream combination can be constructed, which corresponds to four modes: R1, R2a R2b and R3. All modes and respective sub-bitstream combinations are given in Table 1.

Table 1 Sub-bitstream combination for each mode

Mode	Sampl-ing rate [kHz]	G.711 core layer (64 kbit/s)	Lower-band enh. layer (16 kbit/s)	Higher-band enh. layer (16 kbit/s)	Overall bit-rate [kbit/s]
R1	8	X	-	-	64.0
R2a	8	X	X	-	80.0
R2b	16	X	-	X	80.0
R3	16	X	X	X	96.0

The codec has a very simple structure to achieve high quality speech with a low complexity. The proposed codec is deliberately designed without any inter-frame predictions, to avoid annoying artifacts when the switching of the enhancement layers, which is required for the partial mixing in wideband MCU operations. This design policy has another advantage that it also improves FER performance. Gain and polarity of each enhancement layer are quantized separately from other parameters, thus level controlling can easily be exercised without fully decoding and re-encoding of the enhancement layers, i.e. gain code can be just replaced with another code to control the signal level.

The total of analysis and synthesis delays caused by the split-band QMF is 1.875 ms, and the delay caused by the higher-band enhancement layer is 5 ms. Since lower-band codecs do not have delays, the overall algorithmic delay becomes 11.875 ms, including frame processing length (5 ms).

3.2 Core sub-codec

While G.711 is widely regarded to have a sufficient quality for telephone applications, this is not the case when used as a core layer of a wideband codec, because the codec has been optimized for frequency response of telephone speech. For an ordinary input speech signal, its spectral power is usually concentrated in the low frequency range (0 - 1000 Hz), especially in the voiced segments of speech. The G.711 noise has a white spectrum that is highly correlated with the input signal power and this causes a noticeable degradation in the frequency range at 2000 Hz and above. This is why the core encoder consists of a G.711 encoder and a noise shaping processor as shown in Fig. 3. The reason for having a noise shaping processor is to improve the R2b quality, since the white noise, characteristic of G.711 and especially in A-law, must be reduced in order to meet the requirement for R2b, "better than G.722 at 56 kbit/s". Noise shaping process is denoted as follows.

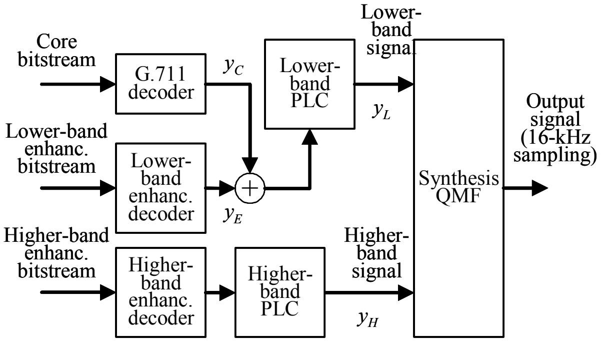
Core bitstream
G.711 decoder
y_C
Lower-band enhanc. bitstream
Lower-band enhanc. decoder
y_E
Lower-band PLC
Lower-band signal
y_L
Higher-band enhanc. bitstream
Higher-band enhanc. decoder
Higher-band PLC
Higher-band signal
y_H
Synthesis QMF
Output signal (16-kHz sampling)

Fig. 2. Decoder block diagram.

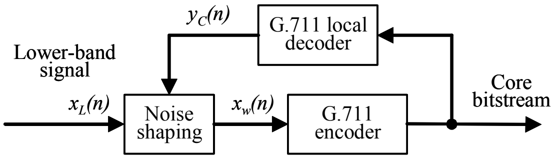

Fig. 3. Core encoder block diagram.

$$x_w[n] = x_L[n] + a(y_C[n-1] - x_L[n-1]) + b, \qquad (1)$$

where $x_w[n]$ is the output of noise shaping module or the input signal for the G.711 encoder at sample n, $x_L[n\text{-}1]$ is the previous input sample of the core encoder, and y_C is the reconstructed signal by G.711. Parameter a is 0.8, and parameter b is 8 for A-law, and 0 for μ-law. This results in quantization noise shaped towards lower frequency, and this is usually masked by speech signal. Since this noise shaping is implemented with a one-tap filter and a fixed coefficient, and it has only a small impact in the computational complexity. The sub-bitstream generated by the core encoder is fully interoperable with the G.711 native decoder and its speech quality is better than the legacy G.711, especially in low-level input signal conditions.

3.3 Lower-band enhancement sub-codec

Figure 4 shows a block diagram of the lower-band enhancement sub-encoder. It reduces the quantization noise of G.711, i.e., the quantization residual signal, $r_L = x_L - y_C$, where x_L is the original input signal in lower-band and y_C is the G.711 decoded signal. A weighted VQ consisting of a shape (notated as c) codebook and a gain (notated as g) codebook are used to achieve this. Here, the inter-frame predictions were not employed, because it would interfere with switching of the sub-bitstreams when partial mixing is applied. The quantizer is intended to minimize the following distance:

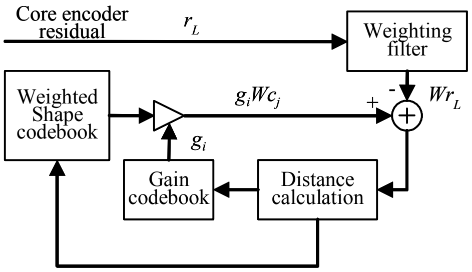

Fig. 4. Lower-band enhancement encoder block diagram.

$$d_L = \left\| W(r_L - g_i c_j) \right\|^2 , \qquad (2)$$

where W, g_i and c_j are a weighting matrix, a scalar gain and a shape vector, respectively. As was used in the core layer, a weighting by a one-tap high-pass filter was used and its transfer function is:

$$H_w(z) = 1 - z^{-1} . \qquad (3)$$

By using a fixed weighting, it is possible to calculate Wc_j offline and store them in a table to avoid convolution for each code vectors c_j. By also keeping the filter tap short, it is possible to perform codebook search by almost the same complexity as the VQ without weighting. Here, the codebook search is done every eight samples. In the decoder, signal is reconstructed by gc_j. The detailed bit-allocation of this sub-codec is shown in Table 3.

Table 3 Bit-allocation for lower-band enhancement sub-codec

Parameter	Bits per 8 sample	Bits per frame
Shape (VQ)	7	35
Polarity (Sign)	1	5
Gain	8	40
Total	16	80
Bit-rate	16.0 kbit/s	

3.4 Higher-band enhancement sub-codec

For the higher-band enhancement sub-codec, an MDCT-based transform coding with *interleaved conjugate-structured VQ* (CS-VQ)is used. Basically, this is a complexity reduced version of a TwinVQ [4] coder. The details of the higher-band encoder are shown in Fig. 5.

In the encoder, the higher-band signal x_H is put through an MDCT filterbank, and MDCT coefficients X_H are obtained. Here, the MDCT shift length is 5 ms. The MDCT coefficients are then normalized using the root mean square (RMS). The normalized coefficients are decimated into a set of 7-sample sub-vectors and those vectors are then independently quantized. Since there are 40 MDCT coefficients in a 5-ms frame, they are divided into 6 sub-vectors by picking up one coefficient every 6 samples. 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, two-channel *conjugate-structured* [5] codebook 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. It should be noted that here, a fast pre-selection algorithm called area-localized method [6] is employed for further complexity reduction. In the pre-selection, 8 candidates are selected among 32 code-vectors in each codebook channels. After pre-selection, the best pair-indices are selected among all combination pairs of pre-selected vectors to minimize the following distance:

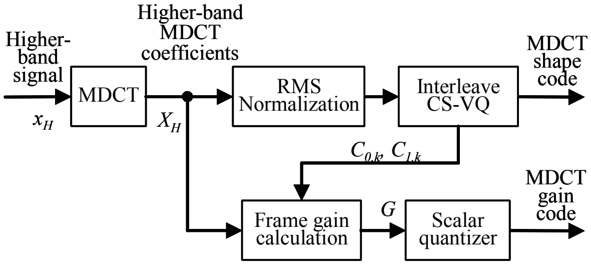

Fig. 5. Higher-band enhancement encoder block diagram.

$$d_k = \left\| X_{H,k} - \frac{C_{0,k} + C_{1,k}}{2} \right\|^2 , \qquad (4)$$

where $X_{H,k}$ is a normalized k-th sub-vector (k = 1,...,6), and $C_{0,k}$ and $C_{1,k}$ are the code-vectors selected from the first and the second codebook channels, respectively. Equation (4) can be re-written as

$$d_k = \left\| X_{H,k} \right\|^2 - X_{H,k} \cdot (C_{0,k} + C_{1,k}) + \frac{\left\| C_{0,k} \right\|^2 + \left\| C_{1,k} \right\|^2 + 2(C_{0,k} \cdot C_{1,k})}{4} , \qquad (5)$$

and complexity is reduced by calculating the power of code-vectors, i.e., $\left\| C_{0,k} \right\|^2$ and $\left\| C_{1,k} \right\|^2$, and their inner products $(C_{0,k} \cdot C_{1,k})$ beforehand and looking-up as table entries.

The frame gain G is calculated as:

$$G = \frac{2 \sum_k X_{H,k} \cdot (C_{0,k} + C_{1,k})}{\sum_k \left\| C_{0,k} + C_{1,k} \right\|^2} . \qquad (6)$$

Then G is compressed using μ-law and is uniformly scalar quantized with 8 bits. The bit-allocation of the coder is shown in Table 4.

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

$$Y_{H,k} = \overline{G} \cdot \frac{C_{0,k} + C_{1,k}}{2} , \qquad (7)$$

where $Y_{H,k}$ is the k-th sub-vector, $\overline{G}$ is the decoded frame gain. All $Y_{H,k}$ 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 y_H.

Table 4 Bit-allocation for higher-band enhancement sub-codec

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	

3.5 Packet-loss (frame erasure) concealment (PLC/FEC)

Lower-band PLC algorithm is similar to the one used in G.711 Appendix I. When a frame erasure is detected, its pitch lag is estimated using the previous frame signals and the lost frame signal is generated by repeating a lag-length waveform at the end of the previous frame. In this implementation, the pitch estimation algorithm was improved by calculating multiple lag candidates and this contributed to reducing the pitch estimation error.

Higher-band PLC algorithm is simple and when a frame erasure is detected, the previous frame buffer data of IMDCT result is copied to the lost frame buffer and is multiplied by an attenuating gain.

4. SUBJECTIVE EVALUATIONS

In order to evaluate the subjective quality of the proposed codec candidate, a subjective listening test was conducted, according to the processing and the quality assessment test plans designed and approved by ITU-T Q7/12 [7]. The listening test consisted of the following five experiments:

Exp 1a: ACR, clean speech, narrowband
Exp 1b: ACR, clean speech, wideband
Exp 2: ACR, music, wideband
Exp 3a: DCR, speech with background noise, narrowband
Exp 3b: DCR, speech with background noise, wideband.

All speech samples used in above experiments were Japanese and 24 subjects required for each experiment are all native Japanese non-experts. All test stimuli were presented to the listeners using Sennheiser HD25, with one capsule removed for monaural listening, in a sound proof room. The listening level was set to the optimum level at the -15 dBPa, which is equivalent to 79 dBSPL, at the ear reference point.

Table 5 gives a summary of the MOS scores and pass/fail judgments of the tested conditions. In this table, "CuT mode" means the test mode of the "coder under test" (i.e., proposed coder), "Reference" means the reference condition of the requirement/objective, "Score$_{\mathrm{CuT}}$" and "Score$_{\mathrm{Ref}}$" are the MOS scores of the proposed coder and the reference coders respectively, R/O indicates whether it is a Requirement or an Objective, and "Pass/fail" shows the final pass/fail judgments. The judgments were made based on the statistical comparison between MOS scores of the proposed codec candidate and the reference codecs, by means of a simple paired t-test at 95% significance level. The candidate codec met all 27 requirements and all 16 objectives. In the qualification phase of the G.711WBE standardization, the same sets of the experiments were also performed by other listening laboratories for cross-checking purpose. Those results were consistent with the previous test results, and the proposed candidate passed all requirements and all objectives except for one Objective, which was high-level input speech quality in R3.

5. COMPLEXITY EVALUATION

Table 6 gives the complexity and required memory of the proposed codec for the speech samples used in the above subjective evaluation. The complexity of the tested candidate codec, which is estimated using basic operators set in the ITU-T Software Tool Library v2.2, is 9.89 WMOPS (weighted million operations per second) in the worst case. This meets the ToR objective ("less than 10 WMOPS"). The memory size of the candidate codec is 1.83 kWords RAM and 3.64 kWords table ROM, and both also met the memory requirements in the ToR.

Table 6 Complexity and memory estimation

		Encoder	Decoder	Total
Complexity (WMOPS)	no FER	7.67	1.21	8.88
	3% FER	7.67	2.22	9.89
Memory (kWords)	RAM	0.76	1.07	1.83
	Table ROM	3.64		

Table 5 Summary of the Subjective Test Results

Exp	CuT mode	Reference	Condition	Score$_{\mathrm{CuT}}$	Score$_{\mathrm{Ref}}$	R/O	P/F
Exp1a	R1	G.711	Clean Speech	3.885	3.677	Req.	Positive Fail*
			3% FER	3.635	3.313	Req.	Pass
Exp3a			Background music	4.719	4.646	Req.	Pass
			Office noise	4.781	4.677	Req.	Pass
			Babble noise	4.740	4.688	Req.	Pass
			Interfering talker	4.656	4.615	Req.	Pass
Exp1a	R2a	16bit PCM	Clean Speech	4.104	4.021	Obj.	Pass
		G.711	3% FER	3.865	3.313	Req.	Pass
Exp1b	R2b	G.722 (56kbit/s)	Clean Speech	4.021	3.458	Req.	Pass
			3% FER**	3.771	2.208	Req.	Pass
Exp2			Music	3.984	3.214	Req.	Pass
Exp3b			Background music	4.542	3.729	Req.	Pass
			Office noise	4.719	3.979	Req.	Pass
			Babble noise	4.698	3.833	Req.	Pass
			Interfering talker	4.531	3.750	Req.	Pass
Exp1b	R3	G.722 (64kbit/s)	Clean Speech	4.146	3.698	Req.	Pass
			3% FER**	3.854	2.260	Req.	Pass
Exp2			Music	3.943	3.234	Req.	Pass
Exp3b			Background music	4.698	3.948	Req.	Pass
			Office noise	4.708	3.958	Req.	Pass
			Babble noise	4.667	3.938	Req.	Pass
			Interfering talker	4.531	3.854	Req.	Pass

* "Positive Fail" means that "Proposed statistically has better quality than reference."
** FER for the reference G.722 was set to 1%.

6. CONCLUSION

As a candidate of the ITU-T G.711WBE, a wideband scalable codec based on UEMCLIP was proposed. The core layer is a G.711 utilized with noise shaping, and has two enhancement layers: a lower band enhancement sub-codec with time-domain weighted VQ and another one for higher band with a interleaved conjugate structure VQ in MDCT domain. While the emphasis in the codec design was on complexity, subjective tests showed that the subjective quality of the proposed codec met all requirements and all objectives specified in the ToR. Complexity evaluation proved that a computational complexity and memory size met the ToR objective and requirement, respectively.

6. REFERENCES

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

[2] 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).

[3] 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, September 2006.

[4] N. Iwakami, T. Moriya and S. Miki, "High quality audio-coding at less than 64 kbit/s by using transform-domain weighted interleave vector quantization TwinVQ", in *Proc ICASSP'95*, pp.3095-3098, 1995.

[5] T. Moriya, "Two-channel conjugate vector quantizer for noisy channel speech coding, " *IEEE JSAC*, vol.10, no.5, pp.866-874, 1992.

[6] N. Iwakami, T Moriya, A. Jin, T. Mori and K. Chikira, "Fast encoding algorithms for MPEG-4 TwinVQ Audio Tool, " in *Proc.ICASSP'01*, pp.3253-3256, 2001.

[7] ITU-T SG16 TD-254/WP3, "G.711 WB extension Qualification Quality Assessment Test Plan", Study Period 2005-2008, Geneva, June 2007 (Source Q.10/16 Rapporteur).