AN 64–80–96 KBIT/S SCALABLE WIDEBAND SPEECH CODING CANDIDATE FOR ITU-T G.711-WB STANDARDIZATION

Balázs Kövesi, Stéphane Ragot, and Alain Le Guyader

France Telecom, R&D Division /TECH/SSTP
2, Av. Pierre Marzin. 22307 Lannion Cedex. FRANCE
Email: {balazs.kovesi,stephane.ragot}@orange-ftgroup.com

ABSTRACT

This article presents a bitrate- and bandwidth-scalable coder submitted as a candidate to the qualification phase of ITU-T embedded G.711 wideband (G.711-WB) standardization. The encoder operates on 5 ms frames and bitrate can be set on a frame basis at either 64, 80 or 96 kbit/s. The input signal, which is by default sampled at 16 kHz, is split in two bands. The low band (0-4000 Hz) is coded by a low-complexity embedded Pulse Code Modulation (PCM) coder at 8-10 bits/sample, including noise feedback at the encoder and PCM-specific post-filtering at the decoder. The high-band (4000-8000 Hz) is coded by a time-domain aliasing cancellation (TDAC) coder derived from ITU-T G.729.1. Frame erasures are concealed using a low-complexity split-band algorithm derived from ITU-T G.722 Appendix IV. The proposed coder passed all requirements of the G.711WB qualification phase. We summarize and discuss formal ITU-T test results.

Index Terms— Speech coding, standardization, ITU, G.711

1. INTRODUCTION

ITU-T SG16 is studying the development of an 64-80-96 kbit/s embedded extension of G.711 with low-delay/low-complexity and wideband (50-7000 Hz) capability to answer the market needs of IP telephony and conferencing applications on high-speed infrastructures, such as optical-fiber access networks (FTTx) [1]. The related work item, called G.711 wideband (G.711-WB), was launched in Q.10/16 in Jan. 2007. The terms of reference (ToR) and time schedule of G.711-WB were finalized in March 2007. Candidate coders were submitted at the end of May 2007. Five organization participated in the qualification phase that ended in July 2007: NTT, France Telecom, VoiceAge, ETRI, and Huawei. Three candidates passed all requirements, one had a marginal failure, and one had 5 failures. The objective of this paper is to present the France Telecom candidate.

The motivations for developing G.711-WB are two-fold. Firstly, G.711 is already widely deployed in narrowband (NB) voice over IP (VoIP) infrastructures. Hence, embedded coding with a core coder interoperable with G.711 is an efficient solution to provide wideband VoIP while keeping backward compatibility with equipments using G.711. Such interoperability constraints already motivated the development of ITU-T G.729.1 in order to smoothly migrate G.729-based equipments to wideband VoIP [2] . Secondly, public switched telephone network (PSTN) is being replaced by broadband IP-based networks. For instance, in the Japanese telecommunication market, fibers to the home (FFTH) are widely used for Internet access lines – the number of the subscribers to FTTH exceeded seven million by end 2006 [1]. In such communication environments bit-rate efficiency becomes far less important than delay and complexity aspects [3].

This paper is organized as follows. Sections 2 and 4 describe the proposed encoder and decoder, respectively. Bit allocation to coding parameters and bistream formats are detailed in Section 3. ITU-T test results are summarized and discussed in Section 5.

2. DESCRIPTION OF THE ENCODER

The encoder takes input signals sampled at 8 or 16 kHz. The input is divided into 5 ms frames. The encoding bit rate can be set to 64, 80 or 96 kbit/s. By default the encoder operates at 96 kbit/s with a 16 kHz-sampled input. Figure 1 presents the encoder architecture.

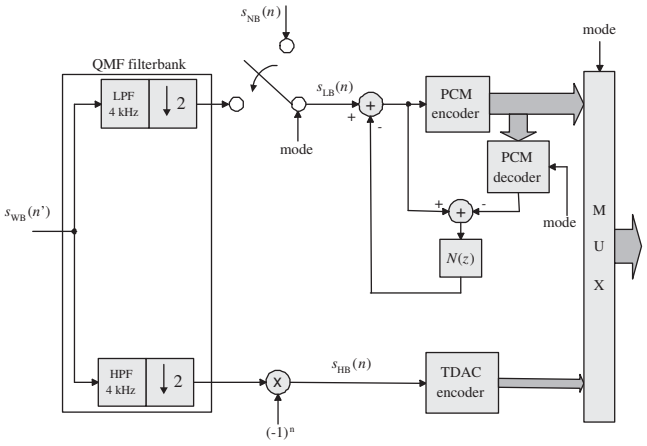

Fig. 1. Block diagram of the encoder.

The input $s_{WB}(n')$ is decomposed into two subbands using a 64-coefficient analysis quadrature mirror filterbank (QMF). This filterbank is the same as in ITU-T G.729.1 [4].

The low band (0-4000 Hz), $s_{LB}(n)$, is coded by an embedded PCM encoder with a bitstream interoperable with ITU-T G.711. Two layers are produced by this embedded PCM encoder: a core layer at 64 kbit/s following the G.711 format and a 16 kbit/s enhancement layer. The core layer contains the 8 bit/sample PCM code according to the G.711 format, with one sign bit, 3 exponent bits identifying a segment and 4 mantissa bits identifying an interval on a given segment [5]. Both A and μ laws are supported. The 16 kbit/s enhancement in low band consists in adding 2 bit/sample to divide each PCM companding subsegment into 4 sub-intervals. Besides, noise feedback based on a local PCM decoder is implemented

to shape the PCM coding noise and improve the quality of the decoded low band. Special attention has been paid when designing this noise feedback to ensure the resulting PCM bitstream can still be decoded by a legacy ITU-T G.711 decoder; hence, no inverse shaping is needed at the decoder.

The high band (4000-8000 Hz), $s_{HB}(n)$, is spectrally folded to restore the natural frequency order and coded in modified discrete cosine transform (MDCT) domain. In this work we modified the time-domain aliasing cancellation (TDAC) encoder of ITU-T G.729.1 [2] for G.711-WB. The MDCT is performed with 5 ms frames (40 samples) and 50% overlap, using sinusoidal windowing. Each MDCT frame consists of 40 transform coefficients. The first 32 coefficients representing the 4000-7200 Hz band are coded, while the last 8 coefficients are discarded. Therefore the effective coded bandwidth is 0-7200 Hz, which exceeds by 200 Hz wideband capability. The first 32 MDCT coefficients are divided into 3 subbands of 8, 8 and 16 coefficients and coded as in G.729.1 by split gain-shape vector quantization (VQ). Each subband gain is quantized using 5 bits with 3-dB-step scalar quantization. The shape is coded using the variable-rate and -dimension MDCT vector quantizer of G.729.1. Given the high bitrate of TDAC coding in G.711-WB (16 kbit/s), the bit allocation to the three MDCT subbands is fixed to minimize complexity. These subbands receive 16, 16 and 32 bits, respectively.

In the following we detail the two innovative parts of the encoder.

2.1. Embedded PCM coding interoperable with G.711

In G.711, logarithmic compression of magnitude is actually approximated by piecewise linear functions. To be specific, the 8 bits of a G.711 codeword specify (in that order): the sign, 3 bits for the segment number of the piecewise linear compression characteristic, and 4 bits for the position on the selected segment.

For an input signal in 16-bit linear PCM format, G.711 actually extracts the 4-bit position by rounding operation, involving a right shift of at least 3 bits. To produce a bitstream scalable coder based on G.711 a simple approach consists in memorizing the two most significant bits normally ignored due to this right shift. It can be done by decomposing the right shift in two parts. In the coder described herein, two bits per sample are saved to form the low-band enhancement layer. The difference in complexity between G.711 and this proposed embedded PCM coder operating at 8-10 bits/sample is negligible, around 0.016 weighted million operations per second (WMOPS). Morever, the extra 2 bits/sample bring +12 dB improvement in signal to noise ratio (SNR).

Note that the generation of the enhancements bits is performed similarly in both A and μ laws.

2.2. Noise feedback coding without side information

Noise feedback coding [6] is used at the low-band encoder, especially to improve quality when the input is coded by G.711-WB and decoded by a legacy G.711 decoder.

It is well-known that at sufficiently high-bit rates (which is the case of G.711 and the proposed embedded PCM coder), PCM coding noise has normally a relatively flat quantization spectrum. For inputs signals with a large spectral dynamic range (around 40 dB or more), this noise may not exceed signal energy. In early predictive speech coders [7, 8, 6], noise shaping techniques have therefore been proposed to make such coding noise inaudible by exploiting the simultaneous masking property. In particular in [7] the shaping filter is derived from a linear-predictive coding (LPC) synthesis filter transmitted to the decoder. However such a filter is not available in G.711 or similar PCM coders. This observation also applied to [9] which extends [7] by including also long-term prediction and vector quantization.

In this work, noise shaping is implemented by a variant of noise feedback [6], as follows. The input $X_{PCM}(z)$ of the PCM encoder is given (in terms of z-transform) by

$$X_{PCM}(z) = X_{LB}(z) - N(z)\left(\hat{X}_1(z) - X_{PCM}(z)\right)$$

where $N(z)$ is a (predictive) shaping filter and $\hat{X}_1(z)$ is the signal decoded at 64 kbit/s. To minimize complexity we used a fixed 1^{st}-order filter for $N(z)$. This is equivalent to applying a fixed "tilt" to the PCM coding noise. Note that more elaborate (adaptive) variants of $N(z)$ can be implemented (e.g. based on LPC and pitch predictors estimated at the encoder). In any case, unlike [7, 6, 9], this form of noise shaping requires no side information.

3. CODING PARAMETERS AND BITSTREAM FORMATS

The encoder generates a scalable bitstream with three hierarchical layers. The core layer (Layer 1) follows the G.711 format. The low-band enhancement layer (Layer 2a) contains the PCM enhancement bits, while the high-band enhancement layer (Layer 2b) contains the TDAC parameters. The bit allocation and layer format are defined in Table 1. Note that there is one unused bit in layer 2b which gives a total usage of 959 bits per 5 ms frames at 96 kbit/s.

Table 1. Bitstream structure for a given 5 ms frame.

(a) hierarchical bitstream structure (embedded layers)

Mode	Layers	Bandwidth	Bit rate (kbit/s)
R1	1	NB	64
R2a	1 + 2a	NB	80
R2b	1 + 2b	WB	80
R3	1 + 2a + 2b	WB	96

(b) detailed bitstream syntax (at 96 kbit/s)

Layer	Parameters	Number of bits
1	PCM bits (G.711 format)	8 bits/sample $\times$ 40 = 320
2a	PCM enhancement bits	2 bits/sample $\times$ 40 = 80
2b	Scale factors of high band	5+5+5
	MDCT VQ	16+16+32
	Unused bit	1
	Subtotal	80
Total per 5 ms frame		480

4. DESCRIPTION OF THE DECODER

The decoder is illustrated in Figure 2. It operates depending on:

- the decoding mode which is specified by the instantaneous bitrate (64, 80, 96 kbit/s) and bitstream format (Layers 1, 1+2a, 1+2b or 1+2a+2b), and
- the bad frame indicator (bfi).

In addition the output sampling rate may be 8 or 16 kHz. In the former case, the output corresponds to the reconstructed low band. In the latter case, the reconstructed low and high bands are combined using a synthesis quadrature mirror filterbank (QMF) of 64 coefficients.

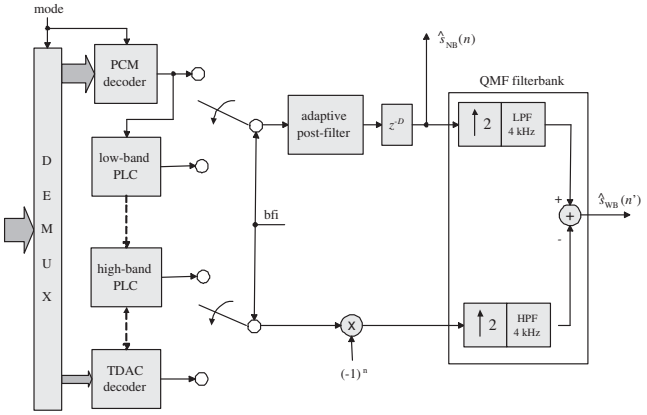

Fig. 2. Block diagram of the decoder.

In the absence of frame erasures (bfi =0), decoding depends on the received mode:

R1 At 64 kbit/s (Layer 1): The core layer is decoded by the PCM decoder and the resulting signal is enhanced by an adaptive symmetric finite-impulse response (FIR) filter of 33 coefficients. The decoded low-band is then time aligned with the high band by a delay unit z-D with D = 24 samples. The high-band is simply decoded to zero.

R2a At 80 kbit/s (Layers 1+2a): Decoding is similar to mode R1. However, the extra 2 bits/sample (PCM enhancement bits) from Layer 2a are used by the PCM decoder to improve the signal to quantization noise ratio (SQNR) in low band. The post-filter also takes into account that the PCM code has 10 bits/sample instead of 8 bits/sample. The high-band is decoded to zero (as in mode R1).

R2b At 80 kbit/s (Layers 1+2b): Low-band decoding is identical to R1. In high band, the TDAC decoder generates a reconstructed signal which is then spectrally folded prior to QMF synthesis.

R3 At 96 kbit/s (Layers 1+2a+2b): Low-band decoding is performed as in mode R2a. High-band decoding is identical to mode R2b.

In case of frame erasure (bfi = 1), a concealment algorithm derived from ITU-T G.722 Appendix IV [10] is used. Indeed, G.722 App. IV has been slightly modified, in particular to operate on 5 ms frames. The high-band PLC algorithm operates in time-domain and updates the modified discrete cosine transform (MDCT) memories of the TDAC decoder to set correct states for the next good frame.

In the following we detail the two innovative parts of the decoder.

4.1. Embedded PCM decoder

The embedded PCM decoder is near-identical to G.711 decoding. The only difference is that, when received, the PCM enhancement bits are simply concatenated to the 4 position bits specified by G.711, and the related rounding value is adapted (divided by 4).

4.2. PCM-specific post-filtering

Post-processing is used in the low-band decoder to improve quality when a narrowband input is coded by legacy G.711 and decoded by G.711-WB. Generally speaking, post-processing is popular, especially in code-excited linear predictive (CELP) speech coding, to reduce coding noise. Two types of post-processing might be applied:

1. Perceptual post-filtering, as in CELP decoding, which exploits available parameters such as linear predicting coding (LPC) coefficients or pitch. Such parameters are not available in G.711-WB for the current frame. They would need to be computed at the cost of extra complexity. Moreover, such postfilters are designed for speech, and they may degrade signals such as music.
2. Conventional noise reduction, which is designed to reduce ambiant noise mixed with speech at the sound capture. However in speech coding, it may be desirable to reproduce with high fidelity the ambiant noise.

In this work we used an alternative approach. Indeed, the coding noise level introduced by PCM can be estimated *a priori* based on the instantaneous magnitude value. Therefore conventional noise reduction may be modified to take into account this *a priori* information. More details on this technique can be found in [11]. This post-processor is implemented using an adaptive FIR filter of 33 coefficients, which results in a delay of 16 samples (2 ms).

In addition to PCM noise reduction, the post-filter includes also a "limiter" that forces each sample at the output of the noise reductor to be in the same A or μ law interval as the (delayed) input of the noise reductor. This additional stage acts as a safety net and avoids possible distortion of reconstructed signals (e.g. noisy speech).

5. QUALITY, DELAY AND COMPLEXITY

5.1. Delay and complexity

The maximal algorithmic delay of proposed coder is 13.9375 ms. The contributions to this delay are detailed below:

- MDCT analysis (current and future frame): 80 samples at 8 kHz $\rightarrow$ 10 ms
- QMF filterbank delay: 63 samples at 16 kHz $\rightarrow$ 3.9375 ms

The high-band lookahead exceeds the delay of 2 ms introduced by the PCM-specific post-processing, which explains why postprocessing does not bring any additional delay. Note that delay goes down to 7 ms if both input and output are sampled at 8 kHz (5 ms for frame length and 2 ms for low-band postprocessing).

The coder submitted to the G.711-WB qualification phase used a mixed implementation with some modules in fixed-point and others in floating-point. After submission, the floating-point to fixed-point conversion has been finalized and the resulting complexity is given in Table 2. It can be noted that the complexity at 64 kbit/s and 80 kbit/s (R1 and R2a modes) is identical which can be beneficial in conferencing applications. Indeed conventional PCM mixing can still be used when receiving both the 64 kbit/s core layer (Layer 1) and the 16 kbit/s narrowband enhancement (Layer 2a).

5.2. Subjective quality

The G.711-WB qualification tests were conducted with ITU-T G.711 A law. Formal quality assessment was organized in 3 main experiments. Exp. 1a tested quality under clean speech conditions in narrowband, while Exp. 1b tested clean speech quality in wideband. Exp. 2 evaluated coder quality under music conditions in wideband. In Exp. 3a (narrowband) and 3b (wideband), noisy speech conditions were tested in the presence of different types of background

Table 2. Fixed-point complexity of the proposed candidate.

		Encoder	Decoder	Codec
Complexity (WMOPS)	R1	–	7.9	11.1
	R2a	–	7.9	11.1
	R2b	–	9.0	12.2
	R3	3.2	9.0	12.2
Static data RAM (kwords)		0.1	0.9	1.0
Dynamic data RAM (kwords)		0.5	1.3	1.8
Data ROM (kwords)		2.8		
Program ROM		2394 basic ops/routines		

noise (background music, office noise, babble noise and interfering talker). Exp. 2, 3a and 3b were performed under error-free condition at -26 dB_{ov}. The effects of input levels variation (-16 dB_{ov}, -26 dB_{ov}, -36 dB_{ov}) and random frame erasures were tested in Exp. 1a and 1b.

Exp. 1a, 1b, and 2 used the Absolute Category Rating (ACR) methodology described in ITU-T P.800. Exp. 3a and 3b used the Degradation Category Rating (DCR) methodology, also described in P.800. All experiments employed 24 subjects, using monaural listening. Speech experiments used 4 talkers (two males, two females) with three sentence-pairs per talker (8s duration). The music experiment (Exp. 2) used 4 genres with three per genre (around 15s duration). All experiments were conducted twice with one home-made test performed by the candidate under test (CuT) and one crosscheck test.

We present here only the CuT-conducted experiments. Speech conditions used the French language. The results of Exp. 1a and 1b (NB and WB clean speech) and Exp. 2 (music) can be found in Figure 3. For the sake of clarity only nominal level (-26 dB_{ov}) is considered. These results show that the proposed coder in R1 mode (at 64 kbit/s) has better speech quality than G.711, which can be attributed to both noise shaping at the encoder and post-processing at the decoder. The 16 kbit/s narrowband enhancement layer (Layer 2a) improves further quality in a more limited amount. As for wideband conditions, the proposed coder has very good speech and music quality at 80 and 96 kbit/s. Note that the use of an MDCT coder in high-band is well suited to code music as well.

The complete set of results and global analysis of experiments can be found in [12]. In particular, it is concluded that the candidate coder described herein passed all quality requirements of the G.711-WB qualification phase [12]. Similar quality results would be obtained with the μ law of ITU-T G.711.

Fig. 3. ITU-T subjective test results for the "Coder under Test" (CuT).

(a) MOS scores of Exp. 1a (NB clean speech)

Coder	FER	CuT test
G.711	0%	3.49
G.711 App.I	3%	3.47
CuT@64k (R1)	0%	4.20
CuT@80k (R2a)	0%	4.55
CuT@64k (R1)	3%	4.12
CuT@80k (R2a)	3%	4.55
G.711 decoded by CuT@64k	0%	3.50

(b) MOS scores of Exp. 1b (WB clean speech)

Coder	FER	CuT test
G.722@56k	0%	3.84
G.722@64k	0%	3.81
G.722@56k+PLC0	1%	3.15
G.722@64k+PLC0	1%	3.17
CuT@80k (R2b)	0%	4.21
CuT@96k (R3)	0%	4.41
CuT@80k (R2b)	3%	4.10
CuT@96k (R3)	3%	4.29

(c) MOS scores of Exp. 2 (music)

Coder	CuT test
G.722@56k	3.85
G.722@64k	3.93
CuT@80k (R2b)	4.12
CuT@96k (R3)	4.18

ACKNOWLEDGMENTS

The authors wish to thank their colleagues, Jean-Luc Garcia and Claude Marro, for providing the PCM noise reduction module.

REFERENCES

[1] ITU-T Contribution AC-0701-06 (WP3/16), "Low-delay Wideband Extension to G.711 for IP Phone Services," Source: NTT, Q.10/16 Rapporteur Meeting, Geneva, 16-19 Jan. 2007 (Study Period 2005-2008).

[2] S. Ragot and al., "ITU-T G.729.1: An 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and Voice over IP," Honolulu, HI, USA, Apr. 2007.

[3] Y. Hiwasaki, H. Ohmuro, T. Mori, S. Kurihara, and A. Kataoka, "A G.711 embedded wideband speech coding for VoIP conferences," *IEICE Transactions on Information and Systems*, vol. E89-D, no. 9, pp. 2542–2552, Sept. 2006.

[4] ITU-T Rec. G.729.1, "An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729," May 2006.

[5] ITU-T Rec. G.711, "Pulse coded modulation (PCM) of voice frequencies," Nov. 1988.

[6] N.S. Jayant and P. Noll, "Noise feedback coding," in *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, chapter 7, pp. 351–371. Prentice Hall, Englewood Cliffs, NJ, USA, 1984.

[7] J. Makhoul and M. Berouti, "Adaptive noise spectral shaping and entropy coding in predictive coding of speech," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 27, no. 1, Feb. 1979.

[8] B.S. Atal and M.R. Schroeder, "Predictive coding of speech signal and subjective error criteria," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 27, no. 3, June 1979.

[9] J.H. Chen, "Novel codec structures for noise feedback coding of speech," in *Proc. ICASSP*, May 2006, vol. 1, pp. 681–684.

[10] ITU-T Rec. G.722 Appendix IV, "A low-complexity algorithm for packet loss concealment with G.722," Nov. 2006.

[11] J.-L. Garcia, C. Marro, and B. Kövesi, "Noise reduction of PCM coding noise," *in preparation*.

[12] ITU-T Contribution AH-07-30 (WP1/16), "Global Analysis Lab Report for the G.711-Wideband Qualification Test," Source: Dynastat, Q.7/12 Rapporteur Meeting, Lannion, 18-22 June 2007.