A 8–32 KBIT/S SCALABLE WIDEBAND SPEECH AND AUDIO CODING CANDIDATE FOR ITU-T G729EV STANDARDIZATION

Stéphane Ragot, Balázs Kövesi, David Virette, Romain Trilling, and Dominique Massaloux

France Telecom, R&D Division /TECH/SSTP
2, Av. Pierre Marzin. 22307 Lannion Cedex. FRANCE

ABSTRACT

This paper describes a 8–32 kbit/s scalable speech and audio coder submitted as a candidate for the ITU-T G729-based Embedded Variable bitrate (G729EV) standardization. The coder is built upon a 3-stage coding structure consisting of: narrowband cascade CELP coding at 8 and 12 kbit/s, bandwidth extension based on wideband linear-predictive coding (WB-LPC) at 14 kbit/s, and MDCT coding in a WB-LPC weighted signal domain from 14 to 32 kbit/s. ITU-T test results showed that this coder passed all the requirements of the G729EV qualification phase.

1. INTRODUCTION

ITU-T SG 16 is currently studying the development of an annex of G729, called G729EV (Embedded Variable bitrate). The required features of the G729EV coder are to provide scalability in the form of an embedded coding scheme, with narrowband (NB) to wideband (WB) audio quality in the range of 8–32 kbit/s for conversational services. Scalability is obtained by structuring the bitstream into embedded layers, using the G729 recommendation as a core coder. Interoperability with G729 allows to deploy wideband VoIP while being compatible with existing VoIP equipments. The foreseen applications are packetized voice over wireline networks (VoIP, VoATM, IP phones, etc.) and high-quality audio/video conferencing.

The G729EV standardization was launched in Jan. 2004. The Terms of Reference (ToR) [1] and initial time schedule of G729EV were finalized in Nov. 2004. A qualification phase was first conducted to check whether candidate coders can pass all requirements. Several companies participated to this preliminary phase: ETRI, Samsung, Siemens–Matsushita–Mindspeed, VoiceAge Corp. and France Telecom. ITU-T tests were performed in July 2005. The objective of this paper is to present the coder submitted by France Telecom as a candidate for this qualification phase.

This paper is structured as follows. The algorithmic details are presented first. The encoder is described in Section 2, while the decoder is presented in Section 3. Then ITU-T qualification tests are summarized and discussed in Section 4.

2. DESCRIPTION OF THE ENCODER

A block diagram of the encoder is shown in Figure 1. The encoder operates with a frame length of 20 ms divided into 2 subframes of 10 ms and 4 subsubframes of 5 ms. It generates an embedded bitstream consisting of 3 layers at 8, 12 and 14 kbit/s and 45 fine-grain-granularity layers from 14 kbit/s to 32 kbit/s by steps of 0.4 kbit/s (one byte per 20 ms frame) including the 10 layers (by steps of 2 kbit/s) required by the ToR of G729EV [1].

As shown in Figure 1, the input signal sampled at 16 kHz is first preprocessed by a 50 Hz high-pass elliptic filter (PRE) of order 2. The resulting signal, $s_{wb}(n)$, is downsampled to 8 kHz using a linear-phase FIR filter of order 40 (LPF) with 3 dB cutoff around 3.6 kHz. The resulting narrowband signal, $s_{nb}(n)$, is encoded by a (two-stage) cascade CELP coder operating at 8 and 12 kbit/s, where the first stage corresponds to a modified G729 encoder (bitstream interoperable with G729). The local narrowband synthesis at 12 kbit/s is then upsampled at 16 kHz.

2.1. Modified G729 coding (8 kbit/s core)

The 8 kbit/s core coder is derived from G729 Main Body [2] with the following modifications to minimize phase distortion and reduce complexity:

- The high-pass elliptic preprocessing of G729 is suppressed.
- The fixed codebook search is replaced by a less complex one.
- The postfiltering and postprocessing in G729 are disabled when computing the local synthesis.

2.2. Cascade CELP coding (12 kbit/s layer)

The 12 kbit/s layer is based on the cascade CELP structure introduced in [3]. It consists of adding an extra fixed codebook to the core coder to enhance the CELP excitation. This additional codebook is optimized to represent to the difference between the original LPC excitation and the LPC excitation reconstructed at 8 kbit/s.

The second-stage fixed codebook corresponds to the G729 fixed codebook followed by a shaping filter. Every 5 ms subsubframe is split into 5 tracks and one pulse is selected in tracks 1, 2, 3 and 4/5. The pulses are postprocessed by an optimized high-pass pulse shaping filter. The second-stage fixed codebook search relies on the algorithm used in the core layer.

In each 5 ms subsubframe the selected codeword in the second-stage fixed codebook is scaled with a gain encoded by intra-frame predictive scalar quantization with 3 bits. The prediction corresponds to the first-stage fixed codebook gain in the same subsubframe.

2.3. Bandwidth extension based on WB-LPC (14 kbit/s layer)

The output bandwidth is extended at 14 kbit/s by parametric coding of the higher band. As shown in Figure 1, $s_{wb}(n)$ is preemphasized by $1 - \mu z^{-1}$, where $\mu = 0.68$. A wideband LPC (WB-LPC) analysis of order 18 is performed, and the resulting coefficients are converted to line spectrum frequencies (LSF). These WB-LSF parameters are quantized using 24 bits by a bandwidth-scalable scheme [4, 5]. The WB-LSF quantization method is similar to [5]. However, the decoded NB-LSF parameters from the 8 kbit/s core coder

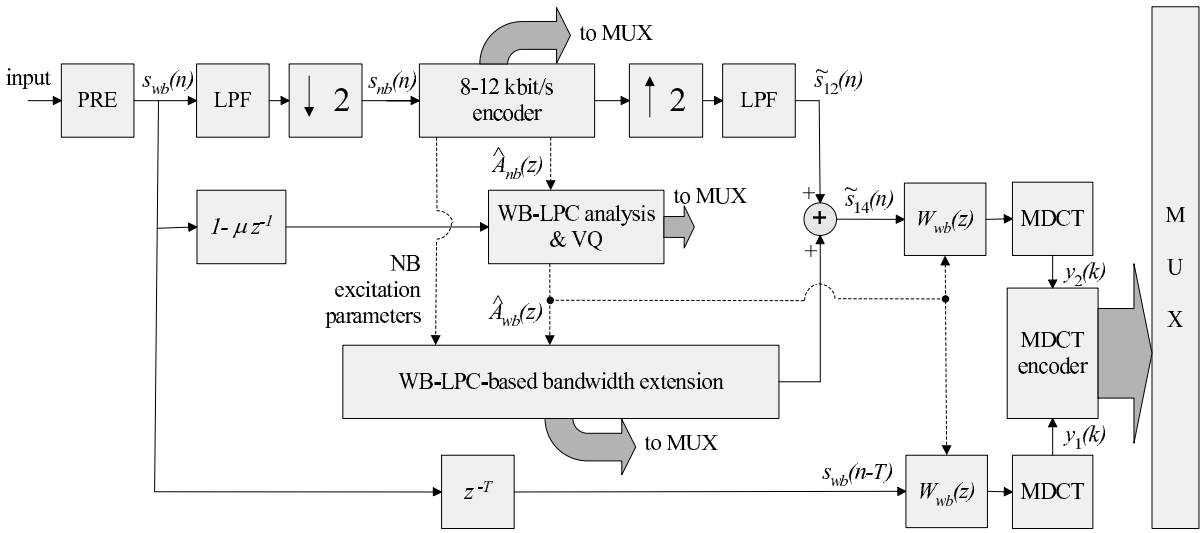

Fig. 1. High-level block diagram of the encoder.

are directly extended and used to predict WB-LSF parameters. The WB-LSF are encoded by predictive 2-stage vector quantization. A 6-bit 18-dimension VQ codebook is used in the first stage, and 17 bits are assigned to 3 split codebooks of the second stage. One bit is used for switching between two predictors. The decoded WB-LSF parameters are interpolated every 5 ms and converted to WB-LPC coefficients, $\hat{A}_{wb}(z)$.

A wideband excitation is generated using the excitation parameters of the 8–12 kbit/s encoder (pitch lag, pitch gain, fixed codewords and codebook gains). This excitation is obtained as in [4] by adding a wideband adaptive excitation and an innovation part produced by upsampling the innovative codevector of the 8-12 kbit/s encoder. The resulting excitation is filtered by $1/\hat{A}_{wb}(z)$ followed by a high-pass filter (HPF) – which is power complementary with the lower-pass decimation filter (LPF) – and scaled every 5 ms by a gain which is quantized using 4 bits. The reconstructed higher band is added to the lower-band local synthesis.

2.4. Predictive MDCT coding (layers above 14 kbit/s)

The delayed signal $s_{wb}(n-T)$ and the local synthesis $\tilde{s}_{14}(n)$ are perceptually weighted by $W_{wb}(z) = \hat{A}_{wb}(z/\gamma)$, where $\gamma = 0.92$. The resulting signals are transformed in frequency domain into $y_1(k)$ and $y_2(k)$ by a modified discrete cosine transform (MDCT) of 640 samples (40 ms). The filter $W_{wb}(z)$ models the short-term frequency masking curve and allows to apply MDCT coding optimized for the mean-square error criterion. It also maps signals into a weighted domain similar to the CELP target domain used at 8 and 12 kbit/s.

The MDCT encoder is a variant of the "TDAC coder" described in [6, 7]. The difference $y(k) = y_1(k) - y_2(k)$ is computed in the 0–3.4 Hz band, while the weighted original signal is taken (i.e. $y(k) = y_1(k)$) in the 3.4–7 kHz band. The effective spectrum therefore consists of a difference signal in the lower band (136 coefficients, 9 subbands) and an original signal in the higher band (144 coefficients, 9 subbands). The other 40 coefficients corresponding to the 7-8 kHz band are discarded.

The effective spectrum is further divided into 18 subbands, where the first subband is composed of 8 coefficients and all the other subbands have 16 coefficients. The spectral envelope corresponding to the subband r.m.s is quantized and Huffman coded. The number of bits allocated to each band is then computed, relying on the decoded spectral envelope. Therefore the bit allocation can be also calculated at the decoder level and no other side information is requested. The MDCT coefficients in each subband are vector quantized, using trained spherical codebooks that are embedded in size and composed of an union of permutation codes. The subbands are transmitted by order of perceptual importance as in [7].

2.5. Bit allocation to coding parameters

The bit allocation is detailed in Table 1.

3. DESCRIPTION OF THE DECODER

The decoder shown in Figure 2 operates depending on the received bit rate R_{dec}.

- If $R_{dec} = 8$ kbit/s, the bitstream is decoded by the G729 Main Body decoder [2] including postfiltering and high-pass elliptic postprocessing (POST). The synthesis is upsampled to 16 kHz.
- If $R_{dec} = 12$ kbit/s, both 8 kbit/s (G729) and 12 kbit/s decoding are performed. The synthesis is also postfiltered and postprocessed. However, the formant postfilter is tuned with the parameters $\gamma_1 = 0.75$ and $\gamma_2 = 0.7$ (instead of $\gamma_1 = 0.7$ and $\gamma_2 = 0.55$). The synthesis is upsampled to 16 kHz.
- If $R_{dec} = 14$ kbit/s, the output of the 8–12 kbit/s decoder is upsampled to 16 kHz without postfilter and postprocessing. The WB-LPC coefficients $\hat{A}_{wb}(z)$ are decoded. A wideband excitation is generated using the same procedure as in the encoder, filtered by $1/\hat{A}_{wb}(z)$, deemphasized, band limited by a high-pass FIR filter (HPF) and scaled by the decoded subsubframe gain. The resulting signal is added to the upsampled narrowband synthesis. To limit the output bandwidth to 0–7 kHz output, the 40 last coefficients corresponding to the 7–8 kHz band are set to zero in MDCT domain. The inverse MDCT is performed and the inverse perceptual weighting filter is applied to obtain the output signal.

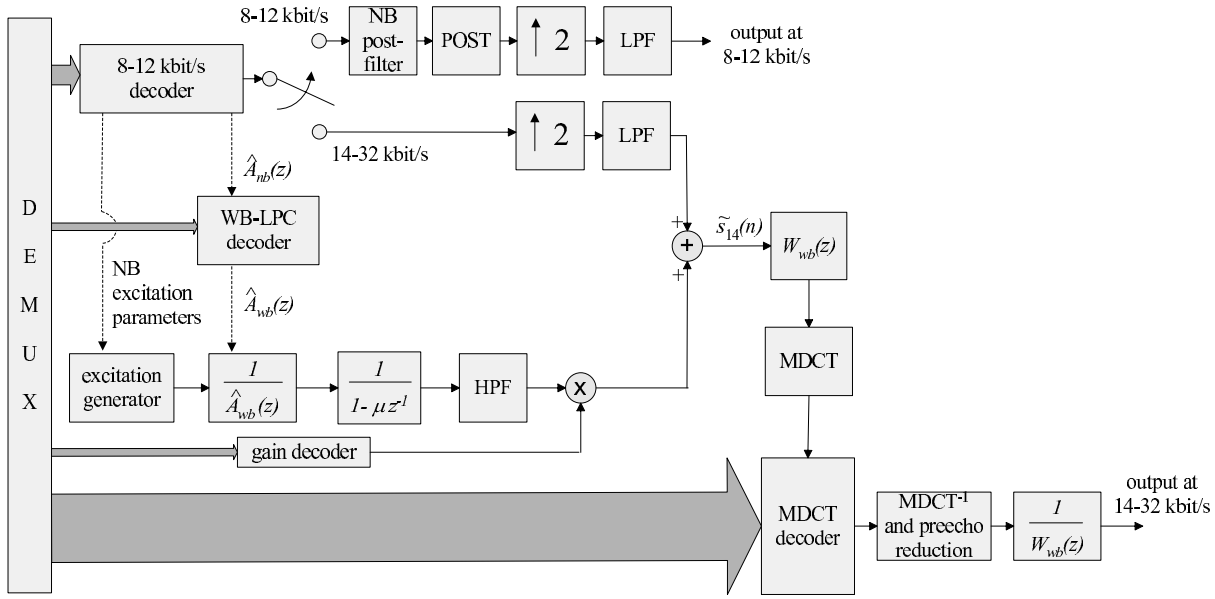

Fig. 2. High-level block diagram of the decoder (in case of received frames).

- For $R_{dec} > 14$ kbit/s, the decoding extends the case $R_{dec} = 14$ kbit/s with an extra step (MDCT decoder) performed in MDCT domain. Depending on the number of extra bits received, the decoder is adapted:

Table 1. Bit allocation.

Parameters	subframe 1 (10 ms)		subframe 2 (10 ms)	
subsubframe (5 ms) →	1	2	3	4
8 kbit/s core (G729-like)				
LSF	18		18	
Pitch lag	8	5	8	5
Pitch parity	1	–	1	–
ACELP codebook	17	17	17	17
Codebook gains	7	7	7	7
Subtotal	80		80	
12 kbit/s layer (cascade CELP)				
ACELP codebook	17	17	17	17
Codebook gain	3	3	3	3
Subtotal	40		40	
14 kbit/s layer (bandwidth extension based on WB-LPC)				
WB-LSF	24			
Gain	4	4	4	4
Subtotal	40			
16–32 kbit/s layers (predictive MDCT coding)				
Scale factors of higher band	Variable			
Scale factors of lower band	Variable			
Normalized MDCT coefficients	Variable			
Subtotal	360			
Total number of bits per frame	640			

(i) If only the spectral envelope is received (partially or fully), the MDCT subbands between 3.4–7 kHz – which correspond to the transformed weighted signal $\tilde{s}_{14}(n)$ – are level-adjusted to gracefully improve quality.

(ii) If the spectral envelope and additional bits are received, the bit allocation is performed in the same manner as in the encoder. The received subbands are decoded by order of perceptual importance. The decoded difference signal between 0–3.4 kHz is added to the transformed weighted synthesis $\tilde{s}_{14}$, while the decoded subbands between 3.4–7 kHz replace their counterpart from the transformed weighted signal $\tilde{s}_{14}$ if their bit allocation is non-zero. For other subbands (non-received or with zero bit allocation), the MDCT coefficients of the transformed weighted synthesis $\tilde{s}_{14}$ are kept. As in the 14 kbit/s layer, the coefficients corresponding to the 7–8 kHz band are set to zero.

3.1. Preecho reduction

The inverse MDCT algorithm involves an inverse FFT and overlap-add of the windowed signal of the previous and current (sine) windows. The current frame corresponds to the last portion of the previous window and the first portion of the current one. Preechos are detected by comparing the energies of the last portion of the previous window and the last portion of the current one. If these energies are significantly different, the shape of the first portion of the current window is modified to reduce preecho.

3.2. Frame erasure concealment

A frame erasure concealment method independent of bitrate is used in case of non-received frames. This method relies on a source-filter model with a wideband LPC filter derived from the WB-LPC filter of the last non-erased frame. If the last non-erased frame was decoded at 8 or 12 kbit/s, the WB-LPC filter is recalculated from the last received samples. An excitation is generated depending on the voicing: in voiced frame, a pitch analysis of the last received

samples is performed around the last pitch lag decoded at 8 kbit/s and is used to extrapolate the excitation ; in unvoiced cases, a non-harmonic component is generated based on the preceding excitation. After LPC synthesis, the energy of the output signal is controlled using a multi-criteria analysis of the past signal components (mainly voiced/unvoiced, pitch and energy slope).

4. SUMMARY OF ITU-T TEST RESULTS

4.1. Algorithmic delay and complexity evaluation

The algorithmic delay is 48.75 ms. The contributions to this delay are 40 ms for the MDCT window, 1.25 ms for the upsampling filter (LPF), 5 ms for G729 lookahead, 1.25 ms for the downsampling filter (LPF) and 1.25 ms for high-pass FIR filter (HPF).

The RAM/ROM and computational complexity evaluations are provided in Table 2. This evaluation is based on actual fixed-point implementation.

Table 2. Estimated storage requirements (in 16-bit words) and computational complexity (in weighted MOPS).

	Encoder (words)	Decoder (words)	Codec (words)
Static RAM	3843	5462	9305
Dynamic RAM	–	–	5000
Tables ROM	–	–	17384
Program ROM	–	–	23000

	Encoder (WMOPS)		Decoder (WMOPS)		Codec (WMOPS)	
8 kbit/s		10.65		3.45		14.10
12 kbit/s	+ 2.3	12.95	+ 0.05	3.5	+ 2.35	16.45
14 kbit/s	+ 8.9	21.95	+ 2.6	6.1	+ 11.5	27.95
32 kbit/s	+ 5.6	27.45	+ 5.4	11.5	+ 11	38.95

4.2. Subjective quality

The subjective quality was evaluated by home-made and cross-checked experiments. Each experiment was conducted with 24 naive listeners using monoaural headphones. The test results showed that the coder submitted by France Telecom passed all subjective quality requirements of the qualification phase [8]. By lack of space only home-made results for conditions at -26 dBov are presented here. The mean opinion scores (MOS) can be found in Table 3. It is worth noting that cross-checked results were consistent with these results.

The submitted coder is better than G729A at 8 kbit/s and equivalent to G729E for clean speech at 12 kbit/s. At 14 kbit/s it is better than G729A and equivalent to G722.2 at 8.85 kbit/s for clean speech. At 24 and 32 kbit/s it is better than G722 at 48 and 56 kit/s (respectively) for speech in various conditions. Furthermore music quality at 32 kbit/s is good (better than G722 at 56 kbit/s).

ACKNOWLEDGMENTS

The authors thank Nicolas Duc for his help in MDCT codebook optimization and complexity evaluation, and Cyril Guillaumé for testing the coder and preparing host labs. They also thank Stéphane Proust for leading the development of the coder.

Table 3. Summary of home-made test results (CuT = Coder under Test). The speech material was in French. For speech with background noise ("Bkgr" conditions), the SNR was respectively 25, 20 and 30 dB in case of music, office and babble background noise.

Bitrate (kbit/s)	Condition	Reference	CuT (MOS)	Reference (MOS)
8	Clean speech	G729A	4.09	3.99
	3% FER	G729A	3.66	3.35
	Music Bkgr	G729A	4.26	4.01
	Office Bkgr	G729A	4.28	4.15
	Babble Bkgr	G729A	4.54	4.49
12	Clean speech	G729E@11.8k	4.35	4.36
	3%FER	G729A	3.87	3.35
	Music Bkgr	G729A	4.55	4.01
	Office Bkgr	G729A	4.63	4.15
	Babble Bkgr	G729A	4.67	4.49
14	Clean speech	G729A	4.06	3.09
	Clean speech	G722.2@8.85k	4.06	4.04
24	Clean speech	G722@48k	4.35	3.99
	1% FER	G722@48k (0%)	4.23	3.99
	Music Bkgr	G722@48k	4.70	4.26
	Office Bkgr	G722@48k	4.65	4.51
	Babble Bkgr	G722@48k	4.77	4.52
32	Clean speech	G722@56k	4.42	4.06
	1% FER	G722@56k (0%)	4.22	4.06
	Music Bkgr	G722@56k	4.79	4.69
	Office Bkgr	G722@56k	4.85	4.60
	Babble Bkgr	G722@56k	4.84	4.72
	Music	G722@56k	4.24	3.98

REFERENCES

[1] ITU-T TD33 (WP3/16), "Terms of Reference (ToR) for the G.729 based Embedded Variable Bit-Rate (G729EV) extension to the ITU-T G.729 Speech Codec," Study Period 2005–2008, Geneva, Nov. 2004 (Source: Q.10/16 Rapporteur).

[2] ITU-T Rec. G729, "Coding of speech at 8 kbit/s using Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP)," March 1996.

[3] R. Drogo De Iacovo and D. Sereno, "Embedded CELP coding for variable bit-rate between 6.4 and 9.6 kbit/s," in *Proc. ICASSP*, April 1991, vol. 1, pp. 681–684.

[4] T. Nomura and al., "A bitrate and bandwidth scalable CELP coder," in *Proc. ICASSP*, May 1998, vol. 1, pp. 341–344.

[5] H. Ehara and al., "Predictive VQ for Bandwidth Scalable LSP Quantization," in *Proc. ICASSP*, March 2005, vol. 1, pp. 137–140.

[6] H. Taddéi and al., "A Scalable Three Bitrate (8, 14.2 and 24 kbit/s) Audio Coder," 107th AES convention, Sept. 1999.

[7] B. Kövesi and al., "A Scalable Speech and Audio Coding Scheme with Continuous Bitrate Flexibility," in *Proc. ICASSP*, May 2004, vol. 1, pp. 273–276.

[8] ITU-T TD71 (WP3/16), "Qualification phase of G729EV: test results (Exp 1-4)," Study Period 2005–2008, Geneva, Jul. 2005 (Source: Q.10/16 and Q.7/12 Rapporteurs).