ITU-T G.729.1: AN 8-32 KBIT/S SCALABLE CODER INTEROPERABLE WITH G.729 FOR WIDEBAND TELEPHONY AND VOICE OVER IP

Stéphane Ragot[1], *Balázs Kövesi*[1], *Romain Trilling*[*], *David Virette*[1], *Nicolas Duc*[*], *Dominique Massaloux*[1], *Stéphane Proust*[1], *Bernd Geiser*[2], *Martin Gartner*[*], *Stefan Schandl*[4], *Hervé Taddei*[3], *Yang Gao*[5], *Eyal Shlomot*[5], *Hiroyuki Ehara*[6], *Koji Yoshida*[6], *Tommy Vaillancourt*[7], *Redwan Salami*[8], *Mi Suk Lee*[9], *and Do Young Kim*[9]

[1]France Telecom R&D, France [2]RWTH Aachen – [3]Siemens Com, Germany – [4]Siemens PSE, Austria
[5]Mindspeed Technologies Inc., USA [6]Matsushita Electric (Panasonic), Japan
[7]University of Sherbrooke – [8]VoiceAge Corp., Canada [9]ETRI, Korea

ABSTRACT

This paper describes the scalable coder – G.729.1 – which has been recently standardized by ITU-T for wideband telephony and voice over IP (VoIP) applications. G.729.1 can operate at 12 different bit rates from 32 down to 8 kbit/s with wideband quality starting at 14kbit/s. This coder is a bitstream interoperable extension of ITU-T G.729 based on three embedded stages: narrowband cascaded CELP coding at 8 and 12 kbit/s, time-domain bandwidth extension (TDBWE) at 14 kbit/s, and split-band MDCT coding with spherical vector quantization (VQ) and pre-echo reduction from 16 to 32 kbit/s. Side information – consisting of signal class, phase, and energy – is transmitted at 12, 14 and 16 kbit/s to improve the resilience and recovery of the decoder in case of frame erasures. The quality, delay, and complexity of G.729.1 are summarized based on ITU-T results.

Index Terms— Speech coding, standardization, ITU

1. INTRODUCTION

ITU-T has recently standardized G.729.1 – a new scalable speech and audio coder developed for the purpose of smoothly migrating narrowband telephony and voice over IP (VoIP) to wideband quality (50–7000 Hz). The related standardization activity was launched in Jan. 2004. At that time, the coder was referred to as G.729-based Embedded Variable bit-rate coding (G.729EV) [1]. Five candidate coders (in floating point) participated in the G.729EV qualification phase which ended in July 2005. The qualified candidates then merged into a single candidate for the optimization / characterization phase of G.729EV. The resulting coder (in fixed point) was approved in May 2006 under the names of ITU-T G.729.1 and Annex J of G.729. The objective of this paper is to give a concise description of G.729.1 and to summarize the associated ITU-T results. A detailed algorithmic description of the coder can be found in [2].

One of the main catalysts of G.729.1 development has been the recent evolution of the telecommunication market towards broadband internet access with integrated services (e.g. triple-play offers including narrowband VoIP, internet and TV). In this context, one differentiating factor is to offer better audio quality using wideband telephony and VoIP. However, such a new service implies significant upgrades of existing infrastructures and terminals. G.729.1 has been designed to answer this market need while allowing a smooth transition from narrowband PSTN quality (300-3400 Hz) to wideband quality (50-7000 Hz). Specifically, the coder is an embedded extension of the widely used ITU-T G.729 speech coding standard.

This paper is organized as follows. Section 2 gives an overview of G.729.1. Its main components are described in Section 3 focusing on the technical innovations brought by G.729.1. Finally, ITU-T test results are summarized and discussed in Section 4.

2. OVERVIEW OF ITU-T G.729.1

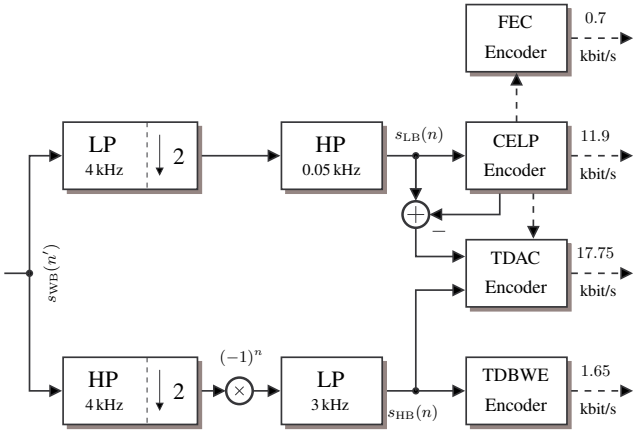
G.729.1 is an 8–32 kbit/s bit-rate and bandwidth scalable speech and audio coder. It supports input and output signals sampled at 8 and 16 kHz. The bitstream comprises 12 embedded layers with a core layer interoperable with ITU-T G.729. The G.729.1 output bandwidth is 50-4000 Hz at 8 and 12 kbit/s and 50-7000 Hz from 14 to 32 kbit/s (per 2 kbit/s steps). G.729.1 operates with a 20 ms "frame length". However, to be consistent with G.729 using 10 ms *frame* and 5 ms *subframes* [3], the 20 ms frames of G.729.1 are referred to as *superframes*.

2.1. Encoder and decoder structure

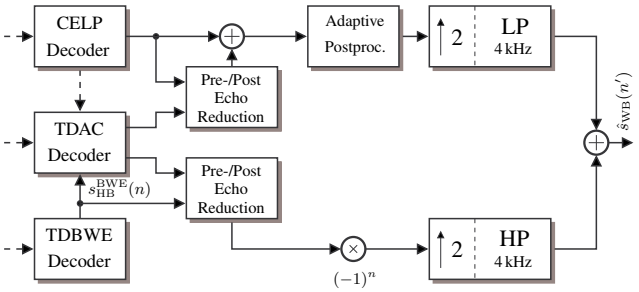
The G.729.1 encoder and decoder are illustrated in Figures 1 (a) and (b), respectively. By default, both input and output signals are sampled at 16 kHz, and the encoder operates at the maximal bit-rate of 32 kbit/s. The input $s_{WB}(n')$ is decomposed into two subbands using a 64-coefficient analysis quadrature mirror filterbank (QMF) [4]. The lower band is pre-processed by an elliptic high-pass filter (HPF) with 50 Hz cutoff and encoded by a cascade (or two-stage) CELP coder. The higher band is spectrally folded, pre-processed by an elliptic low-pass filter (LPF) with 3 kHz cutoff, and encoded by parametric time-domain bandwidth extension (TDBWE). The lower-band CELP difference signal and the higher-band signal $s_{HB}(n)$ are jointly encoded by the so-called time-domain aliasing cancellation (TDAC) encoder which is a transform-based coder. To improve the resilience and recovery of the decoder in case of frame erasures, parameters useful for frame erasure concealment (FEC) are transmitted by the FEC encoder based on available lower-band information.

The decoder operates in an embedded manner depending on the received bit-rate. At 8 and 12 kbit/s the CELP decoder reconstructs a lower-band signal (50-4000 Hz) which is then post-filtered in a way similar to G.729; the result is upsampled to 16 kHz using the QMF synthesis filterbank. At 14 kbit/s, the TDBWE decoder reconstructs a higher-band signal $\hat{s}_{BWE}(n)$ which is combined with the

[*]R. Trilling was with France Telecom R&D, N. Duc worked as a contractor for France Telecom R&D, and Martin Gartner was with Siemens during the development of G.729.1.



(a) encoder



(b) decoder (good frame only)

Fig. 1. Block diagrams of the G.729.1 encoder and decoder.

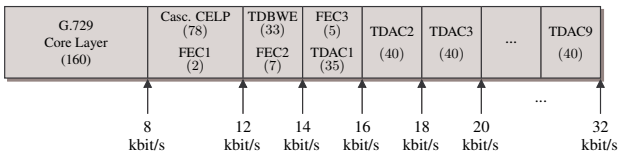
12 kbit/s synthesis in order to extend the output bandwidth to 50-7000 Hz. From 16 to 32 kbit/s, the TDAC decoder reconstructs both a lower-band difference signal and a higher-band signal, which are then post-processed (shaped in time domain) to mitigate pre/post-echo artefacts due to transform coding. The modified TDAC lower-band signal is added to the CELP output, while the modified TDAC higher-band synthesis is used instead of the TDBWE output to improve quality for the whole frequency range.

2.2. Bitstream format

In G.729.1 scalability is obtained by structuring the bitstream into embedded layers, using G.729 as a core coder. The bitstream is divided into 12 embedded layers, as illustrated in Figure 2 (a). The bit rate can therefore be adjusted on-the-fly during a call by simple truncation of the bitstream at any point of the communication chain such as gateways or other devices combining multiple data streams. This highly flexible bit rate adaptation can avoid network congestion and the dropping of packets that severely impair the overall quality. Layer 1 corresponds to a bit-rate of 8 kbit/s and is compliant with G.729 [3]. Layer 2 is a narrowband enhancement layer, while Layers 3 to 12 are wideband enhancement layers. The different parameters multiplexed in each layer are listed in Figure 2 (b); these parameters are defined in more details in Section 3.

2.3. Encoder and decoder modes

G.729.1 includes several modes to complement the default operation mode (called DEFAULT). All modes are listed in Table 1.



(a) hierarchical bitstream structure (embedded layers)

Layer	Parameters	frame 1 (10 ms)		frame 2 (10 ms)	
	subframe (5 ms) →	1	2	3	4
1	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	
2	Tripulse codebook	17	17	17	17
	Codebook gain	3	2	3	2
	Class information (FEC)	–	1	–	1
	Subtotal	40		40	
3	Time envelope mean	5			
	Time envelope split VQ	7+7			
	Frequency envelope split VQ	5+5+4			
	Phase information (FEC)	7			
	Subtotal	40			
4–12	Energy information (FEC)	5			
	MDCT norm shift factor	4			
	Scale factors of higher band	nbits_HB (variable)			
	Scale factors of lower band	nbits_LB (variable)			
	MDCT VQ	nbits_VQ (variable)			
	Subtotal	360			
Total per 20 ms superframe		640			

(b) detailed bitstream syntax

Fig. 2. G.729.1 bitstream structure for a given 20 ms superframe.

The NB_INPUT and NB_OUTPUT modes specify respectively that the input and output are sampled at 8 kHz. The G729_BST and G729B_BST modes offer backward compatibility with G.729/G.729B bitstream format (using 10 ms frames). The LOW_DELAY decoder mode is restricted to 8 and 12 kbit/s and allows to reduce the algorithmic delay of G.729.1 by avoiding the overlap-add operation of the TDAC coding stage. The low-delay mode can be extended to 14 kbit/s – this option is being evaluated by ITU-T.

Table 1. Encoder/decoder operating modes.

Encoder mode	Description	Decoder mode	Description
DEFAULT	16 kHz input	DEFAULT	16 kHz output
NB_INPUT	8 kHz input	NB_OUTPUT	8 kHz output
G729_BST	8 kbit/s only, G.729 bitstream (10 ms)	G729B_BST	decode G.729B frames (10 ms)
–	–	LOW_DELAY	8-12 kbit/s only, low delay (no TDAC)

3. MAIN COMPONENTS OF G.729.1

3.1. Cascade CELP coding (Layers 1 and 2)

The 8 kbit/s core coder is derived from G.729 main body [3] with modifications that reduce phase distortion and complexity. The high-pass elliptic preprocessing of G.729 is omitted, and the G.729 fixed codebook (FCB) search is replaced by a global pulse replacement method [5]. In addition, postfiltering and postprocessing are disabled when computing the local synthesis in G.729.1 encoder. Other modifications are added to improve quality at 8 kbit/s:

- The G.729 open-loop pitch estimation procedure is changed to smooth the pitch track to improve FEC.
- The FCB search is orthogonalized to the adaptive codebook.

The 12 kbit/s layer is based on a cascade CELP structure similar to [1]. It consists of adding an extra FCB to the core coder to enhance the CELP excitation. This additional codebook is optimized to encode the difference between the original LPC excitation and the LPC excitation reconstructed at 8 kbit/s. Specifically, the second-stage FCB is defined by a *tripulse* pattern $-\alpha_{enh}z^{-1} + 1 - \alpha_{enh}z$ with one central pulse +1 and two side pulses. The coefficient $0 \leq \alpha_{enh} \leq 0.34$ is adapted to the voicing of the current subframe with a value 0 for purely voiced segments and 0.34 for purely unvoiced segments. Four tripulses are used per subframe – the associated signs and central locations are indexed with 17 bits in the same way as the G.729 pulse indexing. The 12 kbit/s FCB search uses the 8 kbit/s FCB search with a different target signal and a perceptual filter convolved with $-0.15z + 1 - 0.15z^{-1}$. The 12 kbit/s FCB gain is encoded by intra-frame predictive scalar quantization with 3 bits in subframes with even indices and 2 bits in subframes with odd indices; the prediction is given by the 8 kbit/s FCB gain in the same subframe.

The lower-band adaptive postprocessing in G.729.1 comprises an adaptive postfilter derived from [3] and a high-pass postprocessing (used only at 8 and 12 kbit/s). The γ parameters of the long-term and short-term postfilters depend on the decoder bit rate. Moreover, the adaptive gain control is modified to attenuate fixed-point quantization errors in silence segments (only at 8 and 12 kbit/s).

3.2. Time domain bandwidth extension (Layer 3)

The TDBWE coding stage is derived from the bandwidth extension (BWE) method with side information of [6]. Contrary to classical LPC-based BWE methods, the TDBWE model reconstructs the higher band by shaping an artificial excitation signal according to a desired time envelope (energy per time segments) and a desired frequency envelope (energy per subbands). Time envelope shaping is implemented as a sample-based multiplication by a gain factor, while frequency shaping is performed using a bank of linear-phase finite impulse response (FIR) filters with 2 ms delay. Several improvements have been brought to this model in G.729.1:

- The resolution of the time envelope (2 ms in [6]) is increased so as to better represent sounds like plosives – in G.729.1 the higher-band superframe is divided into 16 segments of 1.25 ms.
- The frequency envelope is computed every 20 ms and interpolated to reduce the number of parameters to be quantized and to get a smooth envelope every 10 ms. The frequency envelope comprises 12 parameters – one per 375 Hz subband.
- A new excitation generator is used based on certain CELP layer parameters. The excitation signal is a weighted mixture of noise and periodic components. The latter are produced by an overlap-add of spectrally shaped and suitably spaced glottal pulses.
- The 28 TDBWE parameters are quantized by mean-removed split VQ with 33 bits. The mean time envelope is coded with 5 bits. The time envelope is quantized in 2 equal blocks with 7 bits, while the frequency envelope is quantized in 3 equal blocks using 5+5+4 bits. Codebooks are trained using a modified K-means algorithm forcing centroids on a rectangular grid.
- The TDBWE synthesis is post-processed by adaptive amplitude compression to remove artefacts ("clicks") due to the lack of coupling between the artificial TDBWE excitation signal and the subsequent shaping.

3.3. Split-band MDCT coding (Layers 4 to 12)

The TDAC coding stage is derived from the transform coding part of [1]. In G.729.1 the lower-band CELP difference signal is filtered per subframe by an "equalized" perceptual filter

$$W_{LB}(z) = \text{fac} \times \hat{A}(z/\gamma_1')/\hat{A}(z/\gamma_2')$$

where $\hat{A}(z)$ is the quantized LPC filter from the 8 kbit/s core coder, $\gamma_1' = 0.96$ and $\gamma_2' = 0.6$; the factor fac is adapted to force a unit gain at the Nyquist frequency, i.e. $W_{LB}(-1) = 1$, which guarantees the spectral continuity between the lower and higher bands. The filter $W_{LB}(z)$ maps the CELP difference signal into a weighted domain similar to the CELP target domain used at 8 and 12 kbit/s.

The modified discrete cosine transform (MDCT) of the (perceptually) weighted CELP difference signal and of the higher-band signal $s_{HB}(n)$ are computed and merged into a full-band spectrum $X(k)$ of 320 coefficients. This spectrum consists of a weighted difference signal in the lower band (160 coefficients) and an original signal in the higher band (120 coefficients); the remaining 40 coefficients, corresponding to the 7-8 kHz band, are discarded. The 280 useful coefficients are divided into 17 subbands of 16 coefficients plus a last subband of 8 coefficients.

The spectral envelope corresponding to the subband root-mean square (r.m.s.) is quantized with 3 dB steps and encoded by a two-mode differential Huffman coder/natural binary coder. The bit allocation per band is then computed, relying on the decoded spectral envelope. Therefore the bit allocation can be also calculated at the decoder and no other side information is required. 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 [1].

The decoding depends on the number of received layers:

- If only the spectral envelope is received (partially or fully), the MDCT subbands between 4–7 kHz – which correspond to the transformed TDBWE synthesis – are level-adjusted to gracefully improve quality.
- If the spectral envelope and additional bits are received, the received subbands are decoded by order of perceptual importance. Subbands that are non-received or with zero bit allocation in the 4–7 kHz range are replaced by the MDCT coefficients of the TDBWE synthesis.

3.4. Pre-/Post-echo reduction (Layers 4 to 12)

The echo reduction mechanism takes advantage of the time-domain coding (CELP and TDBWE) present in the G.729.1. Indeed, the decoded signals from Layers 1 to 3 (from 8 to 14 kbit/s) are free from

pre-/post-echo. Therefore, the related time envelope can be used to reduce the TDAC coding noise (both in lower and higher bands). The echo reduction simply consists in limiting the time envelope of the TDAC synthesis. This limitation is performed only in zones where the echo is detected and disabled in other zones to avoid introducing degradations (e.g. modification of short high-energy segments), especially in the higher band where the time envelope at 14 kbit/s may be inaccurate.

3.5. FEC side information (Layers 2, 3 and 4) and frame erasure concealment

The FEC procedure is derived from the FEC/recovery part of the 3GPP2 VMR-WB speech coder (in generic full-rate encoding type) [7]. At the encoder, 14 bits per superframe are used to send supplementary information, which improves FEC and the recovery of the decoder after frame erasures. The FEC parameters consist of signal classification information (2 bits), phase information (7 bits) and energy (5 bits). They are distributed in Layers 2, 3 and 4 respectively, so at to minimize the impact of bits "stolen" to the cascade CELP, TDBWE and TDAC coding stages.

The FEC follows a split-band approach: in lower band the LPC excitation is reconstructed and filtered by the estimated LPC synthesis filter ; in the higher band, the decoder is supplied with the previously received TDBWE time and frequency envelope parameters – the TDBWE mean-time envelope is attenuated by 3 dB after each erasure. Note that:

- The FEC bits are only used in the lower-band decoding. The class information controls the muting factor. The phase control is used to resynchronize the excitation signal of erased superframes in order to improve the convergence in the correctly received consecutive superframes and reduce error propagation. The energy information is used for energy control of the lower band and compensate for the mismatch between the excitation energy and the gain of the LPC synthesis filter.
- The class information is not available if only the 8 kbit/s core layer is received; in this case, signal classification is performed at the decoder.

In DEFAULT decoder mode, there is a 20 ms delay between the narrowband CELP output and the actual lower-band synthesis – this delay is due to the overlap-add operation of the inverse MDCT needed to generated the lower-band difference signal in the TDAC decoder. This 20 ms delay is equivalent to a lookahead for the FEC/recovery procedure, which allows to perform interpolation of parameters (instead of extrapolation) if only one superframe is erased.

3.6. Bit-rate switching procedure

The G.729.1 coder operates at narrowband bit rates (8 and 12 kbit/s) and wideband bit rates (14 kbit/s and above). Without any appropriate method, a fast switching between these two sets of bit rates results in severe artefacts. Therefore, the decoder includes two mechanisms for bit-rate switching: a cross-fading between the input and output of the lower-band high pass post-processing and a fade-in which forces a slow transition (1 second) from narrowband to wideband.

4. QUALITY, DELAY AND COMPLEXITY

4.1. Delay and complexity

In DEFAULT mode, the algorithmic delay of G.729.1 is 48.9375 ms. The contributions to this delay are 40 ms for the MDCT window (current superframe + lookahead), 5 ms for LPC lookahead, and 3.9375 ms for analysis-synthesis QMF filterbank. Note that for an encoder in NB_INPUT mode and a decoder in NB_OUTPUT and LOW_DELAY mode, the algorithmic delay is reduced to 25 ms.

The RAM/ROM requirements for G.729.1 are the following (in 16-bit words): 5 kword for static RAM, 3.7 kword for dynamic RAM, 8.5 kword for data ROM, and around 32 kword for program ROM. The computational complexity is detailed in Table 2 – this table shows that G.729.1 is also scalable in complexity.

Table 2. Observed worst-case complexity of G.729.1 in DEFAULT mode (in WMOPS using STL2005 v2.1).

Rate (kbit/s)	Encoder		Decoder		Coder	
8		11.65		7.21		18.86
12	+ 2.81	14.46	+ 0.03	7.24	+ 2.86	21.70
14	+ 1.41	15.87	+ 2.54	9.78	+ 3.95	25.65
32	+ 5.57	21.44	+ 4.57	14.35	+ 10.14	35.79

4.2. Subjective quality

The quality was evaluated by formal ITU-T subjective tests in 6 laboratories. Each experiment was conducted in two languages with 32 naive listeners using monaural headphones. The complete results of step 1 of the optimization / characterization phase can be found in [8]; note that step 2 (characterization) is ongoing. G.729.1 is better than G729A at 8 kbit/s and equivalent to G.729E for clean speech at 12 kbit/s. At 14 kbit/s it is better than G.729A at 8 kbit/s and G.722.2 at 8.85 kbit/s for clean speech. At 24 and 32 kbit/s it is better than G.722 at 48 and 56 kbit/s (respectively) for speech in various conditions. Furthermore music quality at 32 kbit/s is good for a conversational coder (equivalent to G.722 at 56 kbit/s).

ACKNOWLEDGMENTS

The authors acknowledge the contributions of Cyril Guillaumé, Peter Jax, Toshiyuki Morii, Bruno Bessette, Guy Richard, and Jongmo Sung in the development of ITU-T G.729.1.

REFERENCES

[1] S. Ragot et al., "A 8–32 kbit/s Scalable Wideband Speech and Audio Coding Candidate for ITU-T G729EV Standardization," in *Proc. ICASSP*, May 2006, vol. 1, pp. 1–4.

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

[3] ITU-T Rec. G.729, "Coding of Speech at 8 kbit/s using Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP)," March 1996.

[4] J. Johnston, "A filter family designed for use in quadrature mirror filter banks," in *Proc. ICASSP*, Apr 1980, vol. 5, pp. 291–294.

[5] E.-D. Lee, M.-S. Lee, and D.-Y. Kim, "Global pulse replacement method for fixed codebook search of ACELP speech codec," in *Proc. IASTED CIIT*, 2003.

[6] P. Jax et al., "An Embedded Scalable Wideband Codec Based on the GSM EFR Codec," in *Proc. ICASSP*, May 2006, vol. 1, pp. 5–8.

[7] 3GPP2 C.S0052-0, "Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB)," June 2004.

[8] ITU-T SG16 Temporary Document, "LS on audio issues," TD 202/GEN, ITU-T Q.10/16, Study Period 2005-2008, Geneva, Switzerland (Source: Rapporteurs Q7/12).