AN EMBEDDED VARIABLE BIT-RATE CODER BASED ON GSM EFR: EFR-EV

Sung-Kyo Jung, Stéphane Ragot, Claude Lamblin, and Stéphane Proust

France Telecom R&D/TECH/SSTP, 22307 Lannion, France

E-mail: {*sungkyo.jung,stephane.ragot,claude.lamblin,stephane.proust*}*@orange-ftgroup.com*

ABSTRACT

This paper describes a 12.2-32 kbps scalable wideband speech and audio coder interoperable with GSM enhanced full-rate (EFR). This coder, referred to as EFR-EV, is designed using the ITU-T G.729.1 multi-stage coding structure. Specifically, EFR-EV consists of three stages: a code-excited linear prediction (CELP) stage derived from EFR, time-domain bandwidth extension (TDBWE), and time-domain aliasing cancellation (TDAC). In this paper, we show that the G.729.1 extension layers (i.e. TDBWE and TDAC) are quite generic for scalable codec design in the sense that they can be applied to EFR with limited adjustments. In addition, we propose a minor modification of the bit allocation procedure in TDAC stage, exploiting spectral masking only for higher frequency bands. The performance of EFR-EV and G.729.1 are evaluated in terms of objective/subjective quality, algorithmic delay, and complexity.

Index Terms— Scalable wideband coder, GSM EFR, embedded coding, EFR-EV, G.729.1

1. INTRODUCTION

Wideband (WB, 0.05-7.0 kHz) speech codecs have been standardized to provide improved audio quality to customers. However, legacy narrowband (NB, 0.3-4.0 kHz) standard codecs are still in use due to codec deployment costs in networks. So, a key issue to WB codec development is interoperability with widely deployed NB codecs. To allow a smooth transition from NB to WB telephony and voice-over IP (VoIP), ITU-T has standardized G.729.1 [1, 2], an embedded coder bitstream interoperable with G.729 [3].

Indeed nowadays different networks use specific incompatible codecs (e.g. EFR and EVRC in GSM and CDMA systems, respectively). Therefore new speech and audio coders are expected to provide not only better quality, but also interoperability with legacy coders that are already widely deployed. Interoperability allows to reduce the cost associated with deploying new coders and smoothly migrate towards new services. After G.729.1 standardization in ITU-T, it has been suggested that a generic use of the embedded variable (EV) structure in G.729.1 can also extend other NB standards while easing transcoding between these G.729.1-based EV coders. In [4], an EV coder based on the G.729.1 embedded structure, EVRC-EV, has been investigated using EVRC at 8.85 kbps as a core coder. In this paper we investigate how the G.729.1 embedded coding structure could be applied to another NB standard CELP coder.

Specifically, we describe the design of an embedded coder with GSM EFR [5] as a core coder. The proposed embedded coder is based on the G.729.1 coding structure. Therefore an overview of ITU-T G.729.1 with its three coding stages is given in Section 2. Section 3 presents an embedded coder bitstream interoperable with EFR. Section 4 proposes a new bit allocation method to improve subjective quality at higher bit rates, especially in music signals. Experimental results are presented and discussed in Section 5, before concluding in Section 6.

2. BACKGROUND: ITU-T G.729.1 CODER

ITU-T G.729.1 coder is an embedded speech and audio coder providing 12 bit rates from 32 kbps down to 8 kbps, with WB rendering at 14 kbps and above. At 32 kbps the bitstream of G.729.1 comprises 12 layers, referred to as Layers 1 to 12. The EV structure allows bit rate adjustment on the fly during a call by simple bitstream truncation at any point of the communication chain. Moreover the 8 kbps core layer provides bitstream interoperability with ITU-T G.729. The G.729.1 codec operates on 20 ms frames and supports input and output signals at 8 and 16 kHz. Its embedded structure consists of three coding stages: CELP, TDBWE, and TDAC stages, as shown in Fig. 1.

The CELP stage encodes the low band (LB, 0.05-4.0 kHz) and produces Layers 1 and 2. It is an embedded CELP coder operating at 8 and 12 kbps [6]. The 8 kbps core coder, derived from G.729, uses the same fixed codebook (FCB). The 12 kbps cascade CELP uses an extra FCB with more high-frequency contribution than the core FCB and emphasizes higher frequencies. The TDBWE stage codes the high band (HB, 4.0-7.0 kHz) and produces Layer 3. The HB signal is modeled by shaping an excitation generated from some CELP parameters with time and frequency envelopes [7]. Hence, WB signals can be generated from 14 kbps. The TDAC stage jointly encodes the LB weighted CELP coding error and the HB input signal by predictive transform coding using modified discrete cosine transform (MDCT) and produces Layers 4 to 12. The MDCT spectrum is divided into 18 subbands, coded by gain-shape vector quantization (VQ). The spectral envelope, computed as the subband root mean square (RMS) in log domain, is quantized then encoded by two-mode lossless coding. The VQ bit allocation is adaptively determined based on the subband perceptual importance, which is derived from the decoded spectral envelope. Then, improved WB quality signals are generated from 16 to 32 kbps by steps of 2 kbps. To improve robustness against frame erasures, encoder-side redundancy is used. Three frame-erasure concealment (FEC) parameters − signal classification information (2 bits), phase information (7 bits) and energy (5 bits) − are multiplexed in Layers 2, 3 and 4 as side information.

3. EV CODER BASED ON GSM EFR: EFR-EV

The G.729.1 coding structure can be easily reused to design an EFR-EV coder by replacing its first coding stage − CELP at 12 kbps − with the 12.2 kbps EFR CELP coder. Hence, EFR-EV has the same

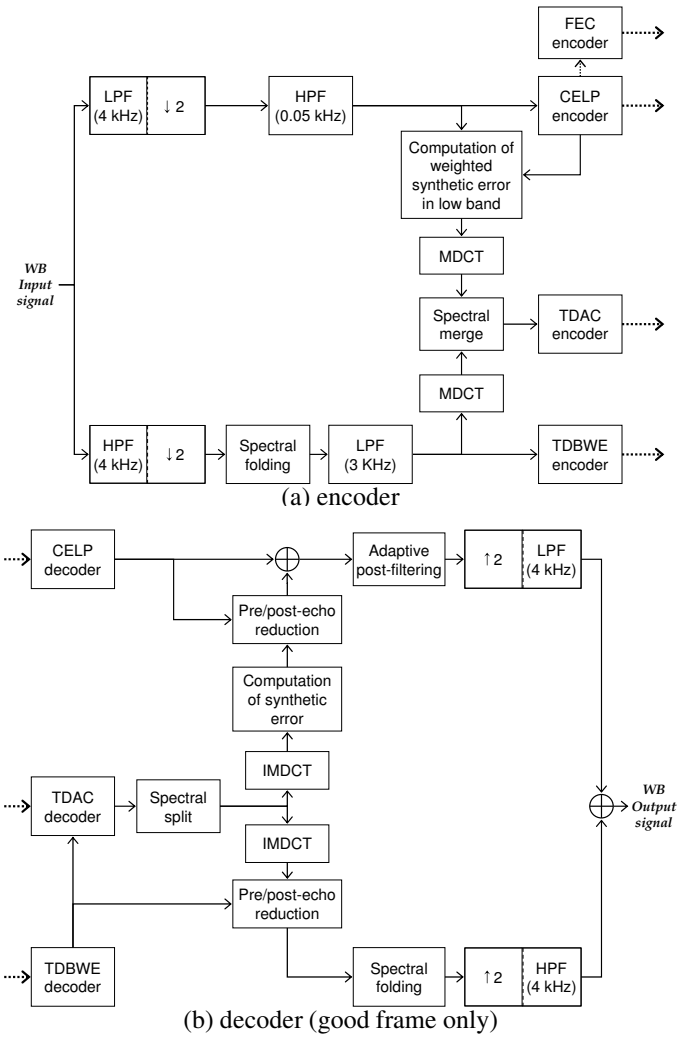

Fig. 1. Block diagrams of encoder and decoder of EFR-EV.

structure as G.729.1 shown in Fig. 1. The adjustments brought to G.729.1 and EFR are described below. Table 1 shows its 11-layered hierarchical bitstream structure with 3 embedded coding stages: CELP, TDBWE, TDAC.

3.1. CELP stage (Layer 1)

The core layer is the GSM 12.2 kbps EFR coder. EFR is implemented using AMR-NB mode at 12.2 kbps, with a look-ahead of 5 ms. The input signal for this stage is filtered by the G.729.1 elliptic high-pass filter with a cut-off frequency of 50 Hz, instead of the EFR high-pass filter with a cut-off frequency of 80 Hz. As in the G.729.1 encoder, some information, such as pitch lag values and energy, used in FEC encoder and TDBWE encoder, is also extracted. The EFR decoder is also modified to use G.729.1 NB post-filter operating at 12 kbps.

Note that the G.729.1 CELP coder and EFR have differences in LPC quantization and excitation signal model. For instance, their pitch models slightly differ: EFR fractional resolution of 1/6 is finer than that of G.729 (1/3), and its lag range wider. Also, the EFR

Table 1. EFR-EV hierarchical bitstream structure

Layer	Parameters	frame (20 ms)			
	subframe (5 ms) →	1	2	3	4
1	2 LSF sets	38			
	Pitch lag	9	6	9	6
	Pitch gain	4	4	4	4
	Algebraic code	35	35	35	35
	Codebook gain	5	5	5	5
	Subtotal	**244**			
2	Time envelope mean	5			
	Time envelope split VQ	7+7			
	Frequency envelope split VQ	5+5+4			
	Class information (FEC)	1		1	
	Phase information (FEC)	7			
	Subtotal	**42**			
3–11	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	**354**			
Total number of bits		**640**			

FCB is a conventional ACELP codebook (10 non-zero pulses with ± 1 amplitude) whereas G.729.1 employs a two-stage codebook with single- and triple-pulse pattern structures. Therefore, the weighted CELP error has different shapes and statistics that impact the input of the enhancement stages.

3.2. TDBWE stage (Layer 2)

The modified EFR core coder is first extended by a TDBWE stage derived from G.729.1. In TDBWE decoder, excitation signal for bandwidth extension is artificially generated based on the encoded parameters of CELP stage. Its shape is determined by the fractional pitch lag. As TDBWE stage in G.729.1 requires a 1/6 fractional resolution two times higher than that of G.729, two methods have been tried to overcome resolution difference between G.729 and EFR pitch lag models: one slightly changes TDBWE module (Eq. (92) in [1]) and directly uses EFR lag, the other (with no changes in TDBWE module) performs a mere mapping of a 1/6 resolution fractional pitch lag to a decimated pitch lag with a 1/3 resolution. Informal listening pair comparison does not show any preference.

3.3. TDAC stage (Layers 3 to 11)

As in G.729.1, the TDAC stage encodes in MDCT domain the full band (FB) signal at the highest bit rate (32 kbps) and generates a 9-layered bitstream with 2 kbps steps. The FB signal is made of the perceptually LB weighted EFR CELP coding error and the HB input signal. To reduce the TDAC coding noise G.729.1 pre-/post-echo cancellation techniques are also applied. The MDCT coefficients are quantized by G.729.1 gain-shape VQ.

As EFR-EV and G.729.1 CELP stages differ, their FB contributions have also different distributions mainly in the LB part. The impact of the CELP stage on the TDAC spectral envelope has been studied. In low-frequency bands (below 1.4 kHz) EFR-EV RMS factors are smaller while they are larger in mid frequency bands around 3.5 kHz. Fig. 2 shows the distribution of the RMS factors for the

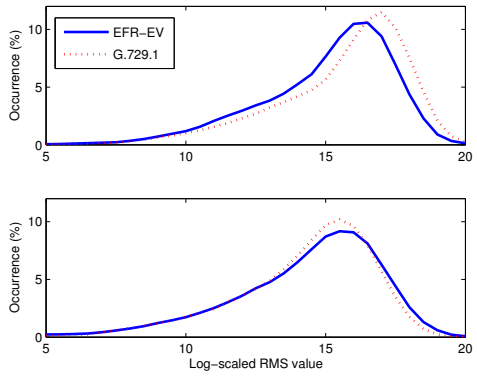

Fig. 2. Histogram of log-scaled RMS factors in subbands 1 to 3 (top) and subbands 8 to 10 (bottom).

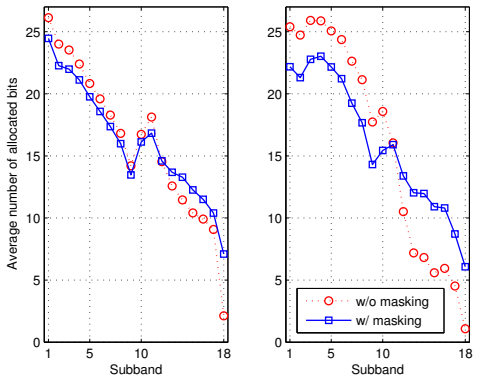

Fig. 3. Average numbers of allocated bits each subband with and without the proposed masking method in speech (left) and music (right) samples.

subbands where the distributions differ the most. Variation in the bit allocation is also entailed as this allocation depends on the decoded spectral envelopes. By comparing average numbers of allocated bits between two coders, we observed that TDAC in G.729.1 gives more bits to the low frequency subbands (1 to 4) while TDAC in EFR-EV focuses more on mid frequency subbands (8 to 12). As TDAC bit allocation depends on the CELP coding error, this indicates that EFR-EV CELP stage provides better description in low band whereas the G.729.1 CELP stage describes better the midband. It was found experimentally that these differences are mainly caused by the FCB. As described in Sec. 3.1, the FCB in G.729.1 and EFR-EV have different structures, and are searched using different criteria. In G.729.1, the second stage FCB with an adaptive triple-pulse pattern has been specially designed to represent high frequency components in a better way than conventional FCB.

3.4. FEC information

As in G.729.1, FEC uses with the same three FEC parameters that are multiplexed in Layers 2 and 3 as side information. The bitstream in Layers 1 and 2 is designed to remain below 14.4 kbps to take into account source codec bit-rate limitation in GSM channel.

4. IMPROVED TDAC CODING

In G.729.1 the TDAC bit allocation is based on the perceptual subband importance and boils down to reverse waterfilling optimization with respect to the mean square error (MSE) criterion [1]. For the MSE criterion to be fully applicable, the FB spectrum should be mapped in an appropriate perceptual domain. However, this spectrum is a concatenation of perceptually weighted (in time domain) EFR coding noise and unweighted HB input. To improve TDAC coding, we propose here to include frequency-domain perceptual weighting of the HB TDAC spectrum. Such a weighting can be efficiently incorporated by modifying the perceptual subband importance $ip(j)$ in the HB originally set to the decoded subband log-energy to a log-scaled *signal-to-mask ratio* while keeping the perceptual importance in LB unchanged.

An approximate masking curve $M(j)$ is computed for each subband of the HB region as the convolution of energy envelope $\hat{\sigma}^2(k)$ and a spreading function $B(\nu)$ as follows [8]:

$$M(j) = \sum_{k=0}^{17} \hat{\sigma}^2(k) \times B(\nu_j - \nu_k), \quad j = 10, \cdots, 17 \qquad (1)$$

where ν_j and ν_k are the center frequencies in *Bark* scale associated to subband indices j and k. The spreading function $B(\nu)$ is defined in log domain as a triangular function with side masking slopes of $+27dB/Bark$ and $-10dB/Bark$ towards higher and lower critical bands, respectively. Hence, the perceptual importance $ip(j)$ is given by:

$$ip(j) = \begin{cases} \frac{1}{2}\log_2\left(\hat{\sigma}^2(j)\right), & j = 0, \cdots, 9 \\ \frac{1}{2}\left[\log_2\left(\frac{\hat{\sigma}^2(j)}{M(j)}\right) + F_{norm}\right], & j = 10, \cdots, 17 \end{cases} \qquad (2)$$

where F_{norm} is a normalization factor to compensate discontinuity between LB and HB in log-scaled energy domain, which is computed by

$$F_{norm} = \log_2\left(\sum_{k=9}^{17} \hat{\sigma}^2(k) \times B(\nu_9 - \nu_k)\right) \qquad (3)$$

Fig. 3 shows average numbers of allocated bits each subband in EFR-EV with and without the proposed masking method for 37,000 speech and music frames. More significant difference in bit allocation has been observed in music samples. Compared to the conventional bit allocation, the proposed method delivers some bits from LB to HB components, especially for music samples. Based on informal subjective tests, the proposed bit allocation method leads better quality in music samples, but no significant improvement in speech samples, with a negligible increase of complexity in both encoder and decoder. When applying the proposed bit allocation method into G.729.1, we can get subjective results similar to EFR-EV.

5. PERFORMANCE EVALUATION

5.1. Objective and subjective quality

The objective performance of EFR-EV and G.729.1 has also been evaluated by WB perceptual evaluation of speech quality (WB-PESQ)

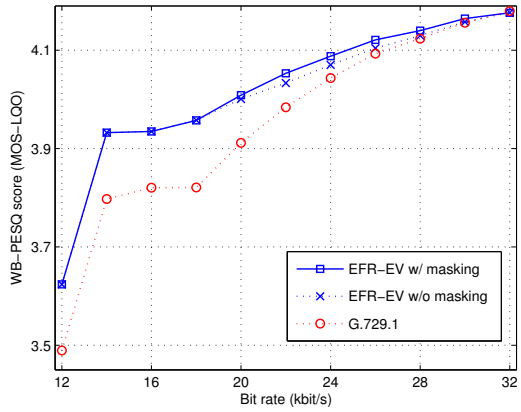

Fig. 4. MOS-LQ0 results over 16 speech sentences.

Table 2. Pair comparison subjective test results

Preference (%)		G.729.1	No preference	EFR-EV
Speech	16 kbps	40.2	23.2	36.6
	24 kbps	34.8	30.4	34.8
	32 kbps	34.3	41.7	24.0
Music	32 kbps	27.9	26.1	46.0

on 16 French sentences from NTT database. Fig 4 shows the average WB-PESQ scores. At low bit rates, EFR-EV scores slightly higher than G.729.1 (WB-PESQ puts more weights on NB description). At higher bit rates the scores are equivalent.

Formal subjective test results with G.729.1 at 12 kbps and EFR are reported in [9]. To compare subjective quality of EFR-EV against G.729.1, three bit rates – 16, 24 and 32 kbps – were used with a data base comprising 8 French sentences with 8 speakers (4 males and 4 females) and 8 music samples. Informal WB pair comparison tests on headsets have been run with 7 expert listeners. The results are summarized in Table 2. EFR-EV provides slightly lower quality or comparable quality to G.729.1 for speech signals, while EFR-EV provides better subjective quality in music samples at the highest bit-rate.

5.2. Algorithmic Delay

As G.729.1 and EFR-EV use the same QMF and MDCT analysis-synthesis, their delays are identical: 48.9375 ms. Both coders have reduced delay mode operations: delay can go down to 25 ms in NB, and to 28.9375 ms for the first WB mode of G.729.1 (14 kbps) and EFR-EV (14.3 kbps).

5.3. Complexity

At 32 kbps, the observed worst-case complexity for G.729.1 codec is 35.8 WMOPS and for EFR-EV 34.4 WMOPS, where WMOPS are evaluated with STL2005 complexity weights [10]. In low delay mode, the complexity is reduced to 25.6 WMOPS for G.729.1 at 14 kbps and 24.3 WMOPS for EFR-EV at 14.3 kbps.

6. CONCLUSION

In this paper we investigated how the G.729.1 embedded coding structure could be applied to other NB CELP coders than G.729 in the view of codec design. An embedded variable bit-rate coder built on top of EFR (EFR EV) has been studied. In addition, we proposed a new bit allocation procedure to improve quality, especially on music signals, at higher bit rates. The overall performance of EFR-EV is comparable to that of G.729.1.

The present work and [4] prove that backward compatibility with existing coders in fixed or wireless network can be maintained while offering enhanced WB quality at higher bit rates. Since the EV coders addressed herein use the same extension layers (TDBWE and TDAC), smart transcoding between these EV coders can be efficiently designed by combining two trancoding techniques: transcoding between NB CELP coders [11] and transcoding between higher layers with compensating the contribution difference of CELP coders. The G.729.1 embedded structure can be therefore expected to offer a generic approach to the enhancement of other existing widely deployed standards with high benefits in terms of deployment efficiency and interoperability cost.

7. ACKNOWLEDGMENT

The authors would like to thank Cyril Guillaumé for his help in implementing the modified TDAC bit allocation.

8. REFERENCES

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

[2] S. Ragot et al., "ITU-T G.729.1: An 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and voce over IP," in *Proc. ICASSP*, May 2007.

[3] ITU-T Rec. G.729, "Coding of speech 8 kbit/s using conjugate-structure albebraic-code-excited linear prediction (CS-ACELP)," Mar. 1996.

[4] ITU-T COM16-C83-E, "Applicability of G.729.1 enhancement layers to 3GPP2 EVRC-family of codecs (Source: Qualcomm Inc.)," Nov. 2006.

[5] 3GPP TS 46.060, "3rd Generation Partnership Project; Technical specification group services and system aspects; Enhanced full rate (EFR) speech transcoding," Jun. 2002.

[6] D. Massaloux et al., "An 8-12 kbit/s embedded CELP coder interoperable with IUT-T G.729 coder: First stage of the new G.729.1 standard," in *Proc. ICASSP*, May 2007.

[7] B. Geiser et al., "Bandwidth extenstion for hierarchical speech and audio coding in ITU-T Rec. G.729.1," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2496–2509, Nov. 2007.

[8] Y. Mahieux et al., "High-quality audio transform coding at 64 kbps," *IEEE Trans. Com.*, vol. 42, no. 11, Nov. 1994.

[9] A. Nagle et al., "Quality impact of diotic versus monaural hearing on processed speech," *AES*, Oct. 2007.

[10] ITU-T Rec. G.191 STL-2005 Manual, "ITU-T software tool library 2005 user's manual," Aug. 2005.

[11] M. Ghenania, "Speech coding format conversion between standardized CELP coders," PhD thesis, Univ. de Rennes 1, France, Jun 2005.