16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

ITU-T EV-VBR: A ROBUST 8-32 KBIT/S SCALABLE CODER FOR ERROR PRONE TELECOMMUNICATIONS CHANNELS

Tommy Vaillancourt [1], Milan Jelínek [1], A. Erdem Ertan [2], Jacek Stachurski [2], Anssi Rämö [3], Lasse Laaksonen [3], Jon Gibbs[4], Udar Mittal[4], Stefan Bruhn[5], Volodya Grancharov[5], Masahiro Oshikiri[6], Hiroyuki Ehara[6], Dejun Zhang[7], Fuwei Ma[7], David Virette[8], Stéphane Ragot[8]

[1] VoiceAge/University of Sherbrooke, [2] Texas Instruments, [3]Nokia, [4] Motorola, [5] Ericsson, [6]Matsushita/Panasonic, [7]Huawei, [8]France Telecom

ABSTRACT

This paper presents ITU-T Embedded Variable Bit-Rate (EV-VBR) codec being standardized by Question 9 of Study Group 16 (Q9/16) as recommendation G.718. The codec provides a scalable solution for compression of 16 kHz sampled speech and audio signals at rates between 8 kbit/s and 32 kbit/s, robust to significant rates of frame erasures or packet losses. It comprises 5 layers where higher layer bitstreams can be discarded without affecting the lower layer decoding. The core layer takes advantage of signal-classification based CELP encoding. The second layer reduces the coding error from the first layer by means of additional pitch contribution and another algebraic codebook. The higher layers encode the weighted error signal from lower layers using MDCT transform coding. Several technologies are used to encode the MDCT coefficients for best performance both for speech and music. The codec performance is demonstrated with selected results from ITU-T Characterization test.

1. INTRODUCTION

In 1999, ITU-T Study Group 16 started to study variable bit rate coding of audio signals. Out of this initial work came Question 9/16, with a goal to standardize a unique "toll-quality" audio embedded codec with wider scope of applications than the coders selected by regional standards bodies. Packetized voice, high quality audio/video conferencing, 3rd generation and future wireless systems (4th generation, WiFi), and multimedia streaming were specified as the primary applications. To cope with heterogeneous access technologies and terminal capabilities, bit-rate and bandwidth scalabilities were also identified as important features of the new codec.

An initial phase was scheduled for March 2007 to select the baseline for further optimization, fixed-point code development, and characterization. This optimization-characterization phase was scheduled for completion in April 2008, to be followed by the standardization of additional super-wideband and stereo extension layers. Four candidate codecs were evaluated in the selection phase. A solution jointly developed by Ericsson, Motorola, Nokia, Texas Instruments and VoiceAge was selected as the baseline codec for further collaboration [1]. Nine other companies declared an intention to participate in the collaboration phase, with four of them contributing technology to the baseline codec and improving its performance, reducing delay, or reducing complexity. These four companies were Matsushita, Huawei, France Telecom and Qualcomm. The description of the resulting codec and summary of its performance are described in the following sections.

The paper is organized as follows. In Section 2 we present a brief summary of the codec features. In Sections 3 and 4, the encoder and the decoder are described. An example of bit allocation is given in Section 5. Finally, a performance evaluation is provided in Section 6.

2. CODEC MAIN FEATURES

The EV-VBR codec is an embedded codec comprising 5 layers; referred to as L1 (core layer) through L5 (the highest extension layer). The lower two layers are based on Code-excited Linear Prediction (CELP) technology. The core layer, derived from the VMR-WB speech coding standard [2], comprises several coding modes optimized for different input signals. The coding error from L1 is encoded with L2, consisting of a modified adaptive codebook and an additional fixed algebraic codebook. The error from L2 is further coded by higher layers (L3-L5) in a transform domain using the modified discrete cosine transform (MDCT). Side information is sent in L3 to enhance frame erasure concealment (FEC). The layering structure is summarized in Table I for the default operation of the codec.

TABLE I : Layer structure for default operation

Layer	Bit-rate	Technique	Internal Sampling rate
L1	8 kbit/s	Classification-based core layer	12.8 kHz
L2	+4 kbit/s	CELP enhancement layer	12.8 kHz
L3*	+4 kbit/s	FEC / MDCT	12.8 / 16 kHz
L4*	+8 kbit/s	MDCT	16 kHz
L5*	+8 kbit/s	MDCT	16 kHz

* Not used for NB input-output

The encoder can accept wideband (WB) or narrowband (NB) signals sampled at either 16 or 8 kHz, respectively. Similarly, the decoder output can be WB or NB, too. Input signals sampled at 16 kHz, but with bandwidth limited to NB, are detected and coding modes optimized for NB inputs are used in this case. The WB rendering is provided for, in all layers. The NB rendering is implemented only for L1 and L2. The input signal is processed using 20 ms frames. Independently of the input signal sampling rate, the L1 and L2 internal sampling frequency is at 12.8 kHz.

The codec delay depends upon the sampling rate of the input and output. For WB input and WB output, the overall algorithmic delay is 42.875 ms. It consists of one 20 ms frame, 1.875 ms delay of input and output re-sampling filters, 10 ms for the encoder look-ahead, 1 ms of post-filtering delay, and 10 ms at the decoder to allow for the overlap-add operation of higher-layer transform coding. For NB input and NB output, the 10 ms decoder delay is used to improve the codec performance for music signals, and in presence of frame errors. The overall algorithmic delay for NB input and NB output is 43.875 ms; 2 ms for the input re-sampling filter, 10 ms for the encoder look-ahead, 1.875 ms for the output re-sampling filter, and 10 ms decoder delay. Note that the 10 ms decoder delay can be avoided for L1 and L2, provided that the decoder is prevented from switching to higher bit rates. In this case the overall delay for WB signals is 32.875 ms and for NB signals 33.875 ms.

The codec is equipped with a discontinuous transmission (DTX) scheme in which the comfort noise generation (CNG) update rate is variable and dependent upon the estimated level of the background noise. An integrated noise reduction scheme [3] can be used if the encoder is limited to L2 during a session.

To satisfy the objective of interoperability with other standards, EV-VBR is equipped with an option to allow it to interoperate with G.722.2 at 12.65 kbit/s. When invoked, the option allows G.722.2 mode 2 (12.65 kbit/s) to replace L1 and L2. Note that this feature makes the codec interoperable also with Mode 2 of the 3GPP AMR-WB standard and Mode 3 of the 3GPP2 VMR-WB standard. The decoder is further able to decode all G.722.2/AMR-WB coding modes. In the G.722.2 interoperability mode, the enhancement layers L3, L4 and L5 are similar to the default operation except that 13 bits less are available in L3 to fit into the 16 kbit/s budget. The addition of the interoperability option has been streamlined due to the fact that the core layer is similar to G.722.2 (operating at 12.8 kHz internal sampling, using the same pre-emphasis and perceptual weighting, etc.)

The encoder-plus-decoder worst case complexity of the fixed point implementation is estimated at around 69 WMOPS using the ITU-T basic operations tool. The worst case complexity of the G.722.2 interoperable option is around 59 WMOPS. The codec memory requirements are 31.8 kWords for ROM and about 25.9 kWords for RAM.

3. ENCODER OVERVIEW

The structural block diagram of the encoder for WB inputs is shown in Figure 1. From the figure it can be seen that while the lower two layers are applied to a pre-emphasized signal sampled at 12.8 kHz as in [4], the upper 3 layers operate at the input signal sampling rate of 16 kHz.

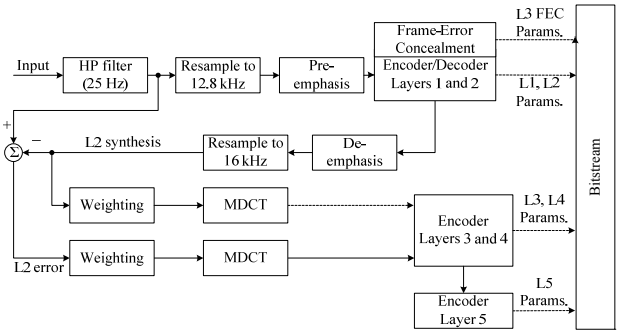

Figure 1: Structural block diagram of the encoder

3.1 Classification based core layer (Layer 1)

To get maximum speech coding performance at 8 kbit/s, the core layer uses signal classification and four distinct coding modes tailored to each class of speech signal; namely Unvoiced coding (UC), Voiced coding (VC), Transition coding (TC) and Generic coding (GC). Some parameters of each coding mode are further optimized separately for NB and WB inputs.

In the core layer, the speech signal is modeled, using a CELP-based paradigm, by an excitation signal passing through a linear prediction (LP) synthesis filter representing the spectral envelope. The LP filter is quantized in the Immitance spectral frequency (ISF) [5] domain using a Safety-Net [6] approach and a multi-stage vector quantization (MSVQ) for the generic and voiced coding modes.

The open-loop (OL) pitch analysis is performed by a pitch-tracking algorithm to ensure a smooth pitch contour, similar to [2]. However, in order to enhance the robustness of the pitch estimation, two concurrent pitch evolution contours are compared and the track that yields the smoother contour is selected.

For NB signals, the pitch estimation is performed using the L2 excitation generated with un-quantized optimal gains. This approach removes the effects of gain quantization and improves pitch-lag estimate across the layers. For WB signals, standard pitch estimation (L1 excitation with quantized gains) is used.

3.1.1 Quantization of LP parameters

To quantize the ISF representation of the LP coefficients, two codebook sets (corresponding to weak and strong prediction) are searched in parallel to find the predictor and the codebook entry that minimize the distortion of the estimated spectral envelope. The main reason for this Safety-Net approach is to reduce the error propagation when frame erasures coincide with segments where the spectral envelope is evolving rapidly. To provide additional error robustness, the weak predictor is sometimes set to zero which results in quantization without prediction. The path without prediction is always chosen when its quantization distortion is sufficiently close to the one with prediction, or when its quantization distortion is small enough to provide transparent coding. In addition, in strongly-predictive codebook search, a sub-optimal codevector is chosen if this does not affect the clean-channel performance but is expected to decrease the error propagation in the presence of frame-erasures. The ISFs of UC and TC frames are further systematically quantized without prediction. For UC frames, sufficient bits are available to allow for very good spectral quantization even without prediction. TC frames are considered too sensitive to frame erasures for prediction to be used, despite a potential reduction in clean channel performance.

There would be too many codebooks if each coding mode and predictor had a unique codebook, and hence some codebooks are reused. Generally speaking, the lower stages of the quantization employ different optimized codebooks to normalize the quantization error. Then common codebooks are used to further refine the quantization.

Two sets of LPC parameters are estimated and encoded per frame in most modes using a 20 ms analysis window, one for the frame-end and one for the mid-frame. Mid-frame ISFs are encoded with an interpolative split VQ with a linear interpolation coefficient being found for each ISF sub-group, so that the difference between the estimated and the interpolated quantized ISFs is minimized.

3.1.2 Excitation coding

The core layer classification starts by evaluating whether the current frame should be coded with the UC mode. The UC mode is designed to encode unvoiced speech frames and, in absence of DTX, most of inactive frames. In UC, the adaptive codebook is not used and the excitation is composed of two vectors selected from a linear Gaussian codebook.

Quasi-periodic segments are encoded with the VC mode, based on the Algebraic CELP (ACELP) technology [4]. VC selection is conditional on a smooth pitch evolution. Given that the pitch evolution is smooth throughout the frame, fewer bits are needed to encode the adaptive codebook contribution and more bits can be allocated to the algebraic codebook than in the GC mode.

The TC mode has been designed to enhance the codec's performance in presence of frame erasures by limiting past frame information usage [7]. To minimize the impact of the TC mode on clean channel performance, it is used only during the most critical frames from a frame erasure point of view – specifically these are frames following voiced onsets. In TC frames, the adaptive codebook in the subframe containing the glottal impulse of the first pitch period is replaced with a fixed codebook of stored glottal shapes. In the preceding subframes, the adaptive codebook is

omitted. In the following subframes, a conventional ACELP codebook is used.

All other frames (in absence of DTX) are processed with the GC mode. This coding mode is basically the same as the generic coding of VMR-WB mode 4 [2] with the exception that fewer bits are available. Thus, one subframe out of four uses a 12-bit algebraic codebook instead of the 20-bit codebook.

The efficiency of the algebraic codebook search has been increased using a joint optimization of the algebraic codebook search together with the computation of the adaptive and algebraic gains by modification of the correlation matrix used in the standard sequential codebook search [8]. A reduced complexity depth-first tree search method [4] is used in GC mode where the number of iterations in the algebraic codebook search is reduced from 4 to 3 with limited SNR loss. To further reduce the complexity of the algebraic codebook search for the critical path, a technology named Path-Choose Pulse Replacement Search (PCPRS) is used in TC and VC frames. This technique is less computationally intensive, but it results in slightly inferior SNR values. Because the encoder complexity for TC and VC frames was higher than for GC frames, using PCPRS technique in those frames was a compromise between better performance and lower worst-case complexity. The PCPRS chooses the best pulse replacement path from two candidate paths in each iteration. These paths have been stored in a table before the actual algebraic codebook search.

To further reduce frame error propagation in the case of frame erasures, gain coding does not use prediction from previous frames in any of the coding modes.

3.2 Second layer encoding (Layer 2)

In L2, the quantization error from the core layer is encoded using an additional algebraic codebook. Further, the encoder modifies the adaptive codebook to include not only the past L1 contribution, but also the past L2 contribution. The adaptive pitch-lag is the same in L1 and L2 to maintain time synchronization between the layers. The adaptive and algebraic codebook gains corresponding to L1 and L2 are then re-optimized to minimize the perceptually weighted coding error. The updated L1 gains and the L2 gains are predictively vector-quantized with respect to the gains already quantized in L1. The output from L2 consists of a synthesized signal encoded in 0-6.4 kHz frequency band. For WB output, the AMR-WB bandwidth extension is used to generate the 6.4-7 kHz bandwidth as in [2].

3.3 Frame erasure concealment side information (Layer 3)

The codec has been designed with emphasis on performance in frame erasure (FE) conditions and several techniques limiting the frame error propagation have been implemented; namely the TC mode, the Safety-Net approach for ISF coding, and the memory-less gain quantization. To further enhance the performance in FE conditions, side information is sent in L3. This side information consists of class information for all coding modes. Previous frame spectral envelope information is also transmitted if the TC mode is used in the core-layer. For other core layer coding modes, phase information and the pitch-synchronous energy of the synthesized signal are sent. The concealment is based on the techniques used in the G.729.1 speech coding standard [9].

3.4 Transform coding of higher layers (Layers 3, 4, 5)

The error resulting from the 2nd stage CELP coding in L2 is further quantized in L3, L4 and L5 using MDCTs. The transform coding is performed at 16 kHz sampling frequency and it is implemented only for WB rendering.

As can be seen from Figure 1, the de-emphasized synthesis from L2 is resampled to a 16 kHz sampling rate. The resulting signal is then subtracted from the high-pass filtered input signal to obtain the error signal which is perceptually weighted and encoded every 20 ms in the transform domain. An asymmetric window, shown in Figure 2, is used to reduce the delay associated to the transform coding stage from 20 to 10 ms while keeping the same number of frequency coefficients. The analysis asymmetric window shape is given by the following equation:

$$w_a(n) = \frac{w_i(n)}{\sqrt{D(n)}}, 0 \le n < 2M \ ,$$

with

$$w_i(n) = \begin{cases} \sin\left[\left(n+\frac{1}{2}\right)\frac{\pi}{(2M-M_z)}\right], & 0 \le n < 2M - M_z \\ 0, & 2M - M_z \le n < 2M \end{cases} .$$

$D(n)$ is defined for $0 \le n < M$ as

$$D(n) = w_i(n)\, w_i(2M\text{-}1\text{-}n) + w_i(n+M)\, w_i(M\text{-}1\text{-}n)$$
$$D(n+M) = D(n),$$

where M=320 denotes the number of MDCT frequency components, and M_z=M/4 is the amount of trailing zeros. The synthesis window is defined as the time reversed analysis window.

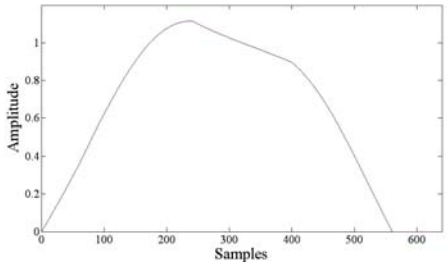

Figure 2 - MDCT analysis window shape.

The MDCT coefficients are quantized differently for speech and music dominant audio contents. The discrimination between speech and music contents is based on an assessment of the CELP model efficiency by comparing the L2 weighted synthesis MDCT components to the corresponding input signal components. For speech dominant content, scalable algebraic vector quantization (AVQ) is used in L3 and L4 with spectral coefficients quantized in 8-dimensional blocks. Global gain is transmitted in L3 and a few bits are used for high-frequency compensation. The remaining L3 and L4 bits are used for the quantization of the MDCT coefficients. The quantization method is the multi-rate lattice VQ (MRLVQ) [10]. A novel multi-level permutation-based algorithm has been used to reduce the complexity and memory cost of the indexing procedure. The rank computation is done in several steps: First, the input vector is decomposed into a sign vector and an absolute-value vector. Second, the absolute-value vector is further decomposed into several levels. The highest-level vector is the original absolute-value vector. Each lower-level vector is obtained by removing the most frequent element from the upper-level vector. The position parameter of each lower-level vector related to its upper-level vector is indexed based on a permutation and combination function. Finally, the index of all the lower-levels and the sign are composed into an output index.

For music dominant content, a band selective shape-gain vector quantization (shape-gain VQ) is used in L3 [11], and an unconstrained pulse position vector quantizer (known as Factorial Pulse Coding, or FPC [12]) is applied to L4. In L3, band selection is performed firstly by computing the energy of the MDCT coefficients. Then the MDCT coefficients in the selected band are quantized using a multi-pulse codebook. A vector quantizer is used to quantize sub-band gains for the MDCT coefficients. For L4, the entire 7 kHz bandwidth is coded using FPC. In the event that the speech model produces unwanted noise due to audio source model mismatch, certain frequencies of the L2 output may be attenuated to allow the MDCT coefficients to be coded more aggressively.

This is done in a closed loop manner by minimizing the squared error between the MDCT of the input signal and that of the coded audio signal through layer L4. The amount of attenuation applied may be up to 6 dB, which is coded using 2 bits. Regardless of which coding method is used in the lower layers, FPC is used exclusively in L5.

4. DECODER OVERVIEW

Figure 3 shows a block diagram of the decoder. In each 20-ms frame, the decoder can receive any of the supported bit rates, from 8 kbit/s up to 32 kbit/s. This means that the decoder operation is conditional on the number of bits, or layers, received in each frame. In Figure 3, we assume WB output, clean channel, and that all layers have been received at the decoder.

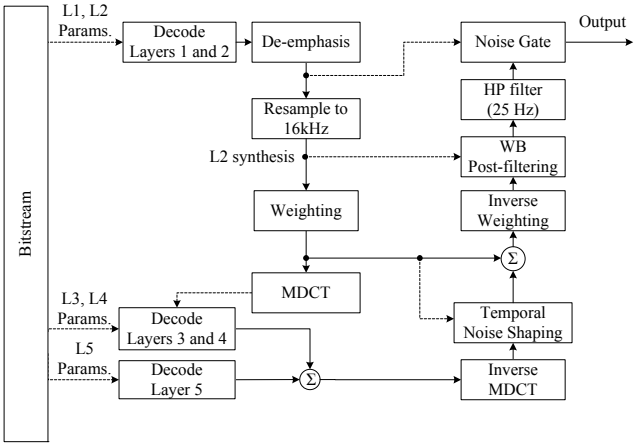

Figure 3 - Block diagram of the decoder (WB, clean channel)

The core layer and the CELP enhancement layer (L1 and L2) are first decoded. The synthesized signal is then de-emphasized and resampled to 16 kHz. After a simple temporal noise shaping, the transform coding enhancement layers are added to the perceptually weighted L2 synthesis. Reverse perceptual weighting is applied to restore the synthesized WB signal, followed by an enhanced pitch post-filter based on [2], a high-pass filter, and a noise gate reducing low-level noise in inactive segments. The post-filter exploits the extra decoder delay introduced for the overlap-add synthesis of the MDCT layers (L3-L5). It combines in an optimal way two pitch post filter signals. One is a high-quality pitch post filter signal of the L1/L2 CELP decoder output that is generated exploiting the extra decoder delay. The other is a low-delay pitch post filter signal of the higher-layer (L3-L5) synthesis signal.

If the decoder is limited to L2 output at call set up, a low-delay mode is used by default, since the additional decoder delay for MDCT overlap-add is not needed. If the decoder output is limited to L1, L2 or L3, a bandwidth extension is further used to generate frequencies between 6.4 and 7 kHz. For L4 or L5 output, the bandwidth extension is not employed and instead the entire spectrum is quantized.

A special feature of the decoder is the advanced anti-swirling technique which efficiently avoids unnaturally sounding synthesis of relatively stationary background noise, such as car noise. This technique reduces power and spectral fluctuations of the excitation signal of the LPC synthesis filter, which in turn also uses smoothed coefficients. As swirling is mainly a problem at low bit rates, it is only activated for L1 signal synthesis (both NB and WB), i.e. if the higher layers are not received. It is based on signal criteria such as voice inactivity and noisiness.

The worst-case complexity of the FE concealment algorithm has been reduced by exploiting the MDCT look-ahead available at the decoder, and distributing the FE concealment algorithm in two consecutive frames.

5. BIT ALLOCATION

Given the fact that the core layer is based on signal classification and several coding modes are used for the core layer, the bit allocation depends to a large extent on the core layer coding mode used. The TC mode has further different bit allocations depending on the position of the first glottal pulse in a frame and the pitch period. If the G.722.2 core-layer option is used, yet another bit allocation is used. An example of the bit allocation for the case when the GC mode is used in the core layer is provided in Table II.

Table II. Example of bit allocation for GC core layer

Layer	Parameter	Subfr. 1	Subfr. 2	Subfr. 3	Subfr. 4
L1	Coding mode	3			
	ISFs	36			
	Energy	3			
	Gains	5	5	5	5
	Adapt. cb.	8	5	8	5
	Algebr. cb.	12	20	20	20
L2	Gains	4	4	4	4
	Algebr. cb.	20	12	20	12
L3	FE param.	16			
	MDCT	62			
L4	MDCT	160			
L5	MDCT	160			

6. PERFORMANCE

The EV-VBR codec was formally evaluated in ITU-T Characterization tests in March 2008. Overall, 9 listening laboratories participated in the tests. The codec was evaluated for 80 reference conditions, each condition evaluated in two different laboratories. Out of these 80 conditions, the codec met the requirements for 78 conditions in both testing laboratories, and for 2 conditions in only one of the two laboratories. The test showed that the most significant progress, with respect to state-of-the-art references, has been made in low bit-rate WB and FE conditions. While not primarily designed for NB inputs, very good performance has been also achieved for NB speech inputs where L1 at 8 kbit/s performed not worse than G.729 Annex E at 11.8 kbit/s for clean speech. Finally, the codec performed very well in noisy conditions both for NB and WB inputs. Selected results extracted from the EV-VBR Characterization test report [13] are summarized below. Results are averaged from both testing laboratories. If not mentioned otherwise, the input level of -26 dBov is assumed.

Figure 4 presents selected MOS results for NB rendering at 8kbit/s and 12 kbit/s at different input levels. The performance is compared to the G.729 and G.729E speech coding standards at 8 kbit/s and 11.8 kbit/s for clean and noisy channel (3% FE rate). The notation LD means that the 10 ms decoder delay was not used. Low bit-rate WB coding performance is demonstrated in Figure 5. The codec performance at 8, 12 and 16 kbit/s is compared to G.722.2 for nominal level clean speech in clean and noisy channel. It can be observed that the codec maintains its performance even in presence of FE rates as high as 8%. 50 Hz random switching among layers has been also tested. Figure 6 shows results for WB rendering for the higher layers (24 and 32 kbit/s) for nominal level clean speech. The conditions tested also included FE conditions where higher erasure rates were applied to higher layers. Figure 7 summarizes the WB performance for music inputs where INT means that L1 and L2 were replaced with G.722.2 interoperable core. Finally, Figure 8 presents results for WB

speech mixed with noise where results are averaged over all noisy conditions (interfering talker, background music, car noise, street noise, babble noise and office noise).

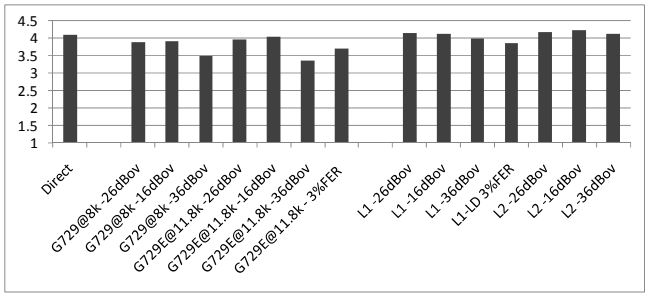

Figure 4 – Performance for NB clean speech

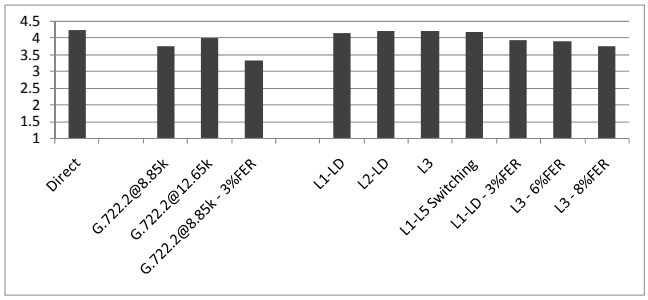

Figure 5 – Performance for WB clean speech at low bit-rates

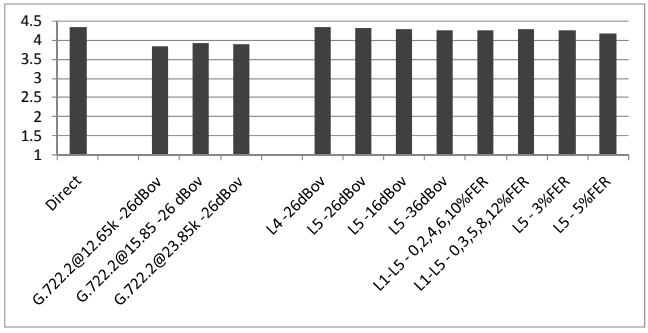

Figure 6 – Performance for WB clean speech at high bit-rates

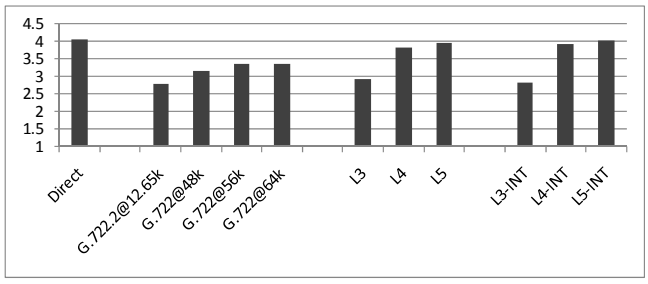

Figure 7 – Performance for WB music

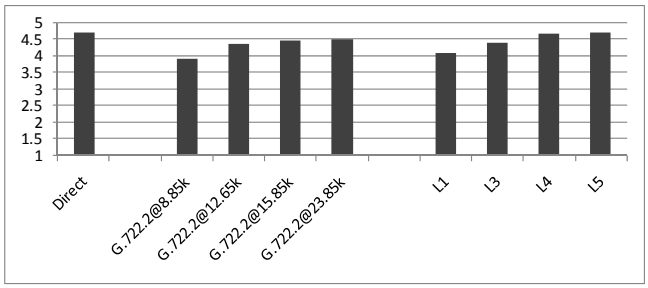

Figure 8 – Performance for WB noisy conditions

7. CONCLUSION

We have presented a new speech and audio embedded codec standardized by ITU-T as recommendation G.718. The structure and main features of the codec were described, and some of the innovative technologies employed have been summarized. Selected results from formal Characterization test show that major advancements with respect to the state of the art references have been achieved in low bit-rate WB and NB speech coding, noisy conditions, and robustness to frame erasures.

ACKNOWLEDGMENT

The authors wish to thank V. Eksler, V. Malenovský, R. Salami, V. Viswanathan, J. Hagqvist, S. C. Greer, J. Svedberg, M. Sehlstedt, E. Norvell, J. P. Ashley, T. Morii, T. Yamanashi, S. Proust, P. Berthet, P. Philippe, B. Kövesi, T. Wang, L. Zhang, P. Huang, and Y. Reznik.

REFERENCES

[1] M. Jelínek, *et al*, "ITU-T G.EV-VBR baseline codec," in *Proc. IEEE ICASSP*, Las Vegas, NV, USA, March, 2008, pp. 4749-4752.

[2] M. Jelínek and R. Salami, "Wideband Speech Coding Advances in VMR-WB standard," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1167-1179, May 2007.

[3] M. Jelínek and R. Salami, "Noise Reduction Method for Wideband Speech Coding," in *Proc. Eusipco*, Vienna, Austria, September 2004, pp. 1959-1962.

[4] B. Bessette, *et al*, "The adaptive multi-rate wideband speech codec (AMR-WB)," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 8, pp. 620-636, November 2002.

[5] Y. Bistritz and S. Pellerm, "Immittance Spectral Pairs (ISP) for speech encoding," in *Proc. IEEE ICASSP*, Minneapolis, MN, USA, April, 1993, vol. 2, pp. 9-12.

[6] T. Eriksson, J. Lindén, and J. Skoglund,, "Interframe LSF Quantization for Noisy Channels," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 5, pp. 495-509, September 1999.

[7] V. Eksler and M. Jelínek, "Transition coding for source controlled CELP codecs," in *Proc. IEEE ICASSP*, Las Vegas, NV, USA, March, 2008, pp. 4001-4004.

[8] U. Mittal, , *et al*, "Joint Optimization of Excitation Parameters in Analysis-by-Synthesis Speech Coders Having Multi-Tap Long Term Predictor," in *Proc. IEEE ICASSP*, Philadelphia, PA, USA, March, 2005, vol. 1, pp. 789-792.

[9] T. Vaillancourt, *et al*, "Efficient Frame Erasure Concealment in Predictive Speech Codecs Using Glottal Pulse Resynchronisation," in *Proc. IEEE ICASSP*, Honolulu, HI, USA, April, 2007, vol. 4, pp. 1113-1116.

[10] S. Ragot, B. Bessette, and R. Lefebvre, "Low-Complexity Multi-Rate Lattice Vector Quantization with Application to Wideband TCX Speech Coding at 32 kbit/s," *Proc. IEEE ICASSP*, Montreal, QC, Canada, May, 2004, vol. 1, pp. 501-504.

[11] M. Oshikiri, *et al*, "An 8-32 kbit/s Scalable Wideband Coder Extended with MDCT-based Bandwidth Extension on top of a 6.8 kbit/s Narrowband CELP Coder," in *Proc. Interspeech*, Antwerp, Belgium, August, 2007, pp.1701-1704.

[12] U. Mittal, J. P. Ashley, and E. Cruz-Zeno, "Low Complexity Factorial Pulse Coding of MDCT Coefficients using Approximation of Combinatorial Functions," in *Proc. IEEE ICASSP*, Honolulu, HI, USA, April, 2007, vol. 1, pp. 289-292.

[13] *Summary of results for G.EV-VBR*, ITU-T Q7/SG12 AH-08-44, Technical Contribution, Lannion, France, April 2008.