ITU-T G.EV-VBR BASELINE CODEC

Milan Jelínek [^], *Tommy Vaillancourt* [^], *A. Erdem Ertan* [#], *Jacek Stachurski* [#], *Anssi Rämö* [*], *Lasse Laaksonen* [*], *Jon Gibbs*[‡], *and Stefan Bruhn*[†]

[^] VoiceAge, [#] Texas Instruments, [*] Nokia, [‡] Motorola, [†] Ericsson,

ABSTRACT

We present the G.EV-VBR winning candidate codec recently selected by Question 9 of Study Group 16 (Q9/16) of ITU-T as a baseline for the development of a scalable solution for wideband speech and audio compression at rates between 8 kb/s and 32 kb/s. The Q9/16 codec is an embedded codec comprising 5 layers where higher layer bitstreams can be discarded without affecting the decoding of the lower layers. The two lower layers are based on the CELP technology where the core layer takes advantage of signal classification based encoding. The higher layers encode the weighted error signal from lower layers using overlap-add transform coding. The codec has been designed with the primary objective of a high-performance wideband speech coding for error-prone telecommunications channels, without compromising the quality for narrowband/wideband speech or wideband music signals. The codec performance is demonstrated with selected test results.

Index Terms— Speech Coding, Audio coding, Embedded Coding, Scalable Coding, ITU

1. INTRODUCTION

In 1999, ITU-T Study Group 16 started to study variable bit rate coding of speech signals. Out of this initial work came Question 9, 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. Among others packetized voice, high quality audio/video conferencing, 3rd generation and future wireless systems (4th generation, WiFi), and multimedia streaming were specified as primary applications. To cope with heterogeneous access technologies and terminal capabilities, bit-rate and bandwidth scalability was another important aspect of the new codec.

A selection phase was scheduled for March 2007 to select the most promising technology among candidate codecs to form the baseline for further optimization, fixed-point code development, and characterization. This optimization-characterization phase is scheduled to end in April 2008, and will be followed by the standardization of additional super-wideband and stereo extension layers.

Four candidate codecs were evaluated in the selection phase. Two candidate solutions performed very similarly and outperformed the other candidates. These two solutions were jointly developed by Ericsson, Motorola, Nokia, Texas Instruments and VoiceAge, and one of them was selected as the baseline for further collaboration. The description of this codec and summary of its performance are given in this contribution.

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 bit allocation is given in Section 5. Finally, a performance evaluation is provided in Section 6.

2. SUMMARY OF CODEC MAIN FEATURES

The Q9 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 the ACELP technology [1]. The core layer, based on 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 an additional innovation codebook. The error from L2 is further coded by higher layers in transform domain using the modified discrete cosine transform (MDCT). The layering structure is summarized in Table I for the default operation of the codec.

TABLE I : Layer structure for default operation

Layer	Bitrate	Technique		Sampling rate	
L1	8 kb/s	Classification-based core layer		12.8 kHz	
L2	+4 kb/s	Algebraic codebook layer		12.8 kHz	
L3*	+4 kb/s	FEC	MDCT	12.8	16 kHz
L4*	+8 kb/s	MDCT		16 kHz	
L5*	+8 kb/s	MDCT		16 kHz	

* Not implemented for NB input-output

The encoder can accept either WB or NB signals sampled at 16 kHz, or NB signals sampled at 8 kHz. Similarly, the decoder output can be WB or NB. 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. Independently of the input signal sampling rate, L1 and L2 internal sampling is at 12.8 kHz. The input signal is processed using 20 ms frames.

The codec delay depends on the sampling rate of the input and output. For WB input and WB output, the overall algorithmic delay is 54.75 ms. It consists of one 20 ms frame, 1.875 ms delay of input and output re-sampling filters, 11.875 ms for the encoder look-ahead, 1 ms of post-filtering delay, and one 20 ms frame delay at the decoder to allow for the overlap-add operation of higher-layer transform coding. For NB input and NB output, the overall algorithmic delay is 55.75 ms. Note that the one-frame transform coding delay is not required for L1 and L2 provided that the decoder is not allowed to switch to higher bit rates. In this case the overall delay is lower by 20 ms both for NB and WB signals.

The codec is equipped with a discontinuous transmission (DTX) scheme with comfort noise generation (CNG) update transmission rate being variable and dependent upon the estimated level of the background noise.

To satisfy the objective of interoperability with other standards, G.EV-VBR is equipped with an option to allow it to interoperate with G.722.2 at 12.65 kb/s. When invoked, the option allows G.722.2 mode 2 (12.65 kb/s) to replace L1 and L2. Note that this feature makes the codec interoperable also with the 3GPP AMR-WB standard and the 3GPP2 VMR-WB standard. The decoder is further able to decode G.722.2 at 8.85 and 6.6 kb/s.

In the G.722.2 interoperability mode, the enhancement layers L3, L4 and L5 are similar to the default operation except that L3 uses fewer bits (to fit into the 16 kb/s budget). The addition of the interoperability option has been streamlined due to the fact that the core ACELP layer is similar to G.722.2 (operating at 12.8 kHz internal sampling, using the same pre-emphasis and perceptual weighting, etc.)

Finally, the codec uses an integrated noise reduction algorithm based on [3] for better estimation of some of the parameters. By default, the denoised signal is not input to the codec for processing. The noise reduction can be however activated through a command line option provided that the communication is limited to L2. This restriction is imposed as the noise reduction is implemented at the internal 12.8 kHz sampling rate and consequently does not perform well for higher layers. In general, noise reduction is however needed only for the lowest bit rates. At higher rates background signals are considered as signals conveying information.

The encoder plus decoder current worst case complexity is estimated at around 57 WMOPS using automated operation counters in floating-point implementation.

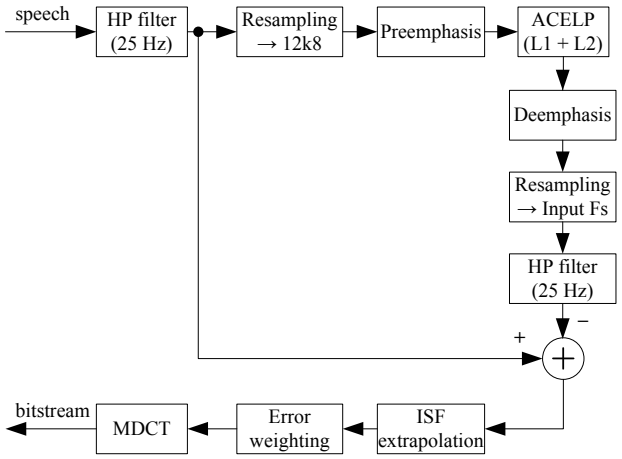

Figure 1: Structural block diagram of the encoder

3. ENCODER OVERVIEW

The structural block diagram of the encoder, for different layers, is shown in Fig. 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 [1], the upper 3 layers operate in the input signal domain sampled at 16 kHz.

The core layer is based on the Code-excited Linear Prediction (CELP) technology where the speech signal is modeled by an excitation signal passed through a linear prediction (LP) synthesis filter representing the spectral envelope. The LP filter is quantized in the Immitance spectral frequency (ISFs) [4] domain using a Safety-Net [5] approach and a multi-stage vector quantization (MSVQ) for the generic and voiced modes. 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 of the Safety-Net approach is to reduce the error propagation due to ISF prediction in case of frame erasures hitting segments where the speech spectral envelope evolves 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 that of a path 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 clean-channel performance but is expected to decrease error propagation in frame-erasures. The ISFs of unvoiced frames and frames following voiced onsets are further systematically quantized without prediction. For unvoiced frames, sufficient bits are available to allow for very good spectral quantization even without prediction. The frames following voiced onsets are too sensitive to frame erasures for prediction to be used and hence it is disabled; despite a potential slight degradation in clean channel conditions.

As there would be too many codebooks if each mode and predictor had a unique codebook, some codebooks are reused. Generally speaking, lower stages 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 twice 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; for each ISF sub-group, a linear interpolation coefficient is found so that the difference between the estimated and the interpolated quantized ISFs is minimized.

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

3.1. Classification based core layer (Layer 1)

To get maximum speech coding performance at 8 kb/s, the core layer uses signal classification and four distinct coding modes tailored for 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.

The frames to be encoded with UC are selected first. UC is designed to encode unvoiced speech frames and, in the absence of DTX, most of the inactive frames. In UC, the adaptive codebook is not used and the excitation is composed of two vectors selected from a linear Gaussian codebook. The excitation gain is coded with a memoryless scalar quantizer.

Quasi-periodic segments are encoded with VC mode. VC selection is conditional on a smooth pitch evolution. It uses ACELP technology, but given that the pitch evolution is smooth throughout the frame, more bits can be attributed to the algebraic codebook than in the GC mode.

The Transition coding mode has been designed to enhance the codec performance in presence of frame erasures by limiting past frame information usage. To minimize at the same time its impact on clean channel performance, it is used only on most critical

frames from a frame erasure point of view – 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 legacy Algebraic CELP (ACELP) codebook is used.

All other frames (in absence of DTX) are processed through a Generic ACELP. This coding mode is basically the same as the generic coding of VMR-WB mode 4 [2] with the exception that less bits are available here. 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 [6].

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

3.2. Second layer ACELP 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. The AMR-WB bandwidth extension is used to generate the missing 6.4-7 kHz bandwidth.

3.3. FE 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 memoryless gain quantization. To further enhance the performance in FE conditions, side information is sent in L3. The side information consists of class information for all coding modes. Previous frame spectral envelope information is further transmitted for the core layer TC. For other core layer coding modes, phase information and the pitch-synchronous energy of the synthesized signal are sent. These parameters help the concealment of erased frames and, more importantly, the recovery of the decoder following erasures. The concealment is similar to the concealment used in the G.729.1 speech coding standard [7].

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 with 50% overlap-add. The transform coding is performed at 16 kHz sampling frequency and it is implemented only for WB rendering.

As can be seen from Fig. 1, the de-emphasized synthesis from L2 is resampled to a 16 kHz sampling rate and high-pass filtered. The resulting signal is then subtracted from the high-pass filtered input signal to obtain the error signal which is weighted and encoded using the MDCT. The MDCT coefficients are quantized using scalable algebraic vector quantization. An MDCT is computed every 20 ms, and its spectral coefficients are quantized in 8-dimensional blocks. An audio cleaner is also applied, derived from the spectrum of the original signal.

The transform coefficients are quantized in the following way. Global gains are transmitted in L3 and a few bits are used for high-frequency compensation. The remaining L3 bits are used for the quantization of the MDCT coefficients. The L4 and L5 bits are used such that the performance is maximized independently at the L4 and L5 levels. The MDCT coefficients are quantized in blocks of 8 bits.

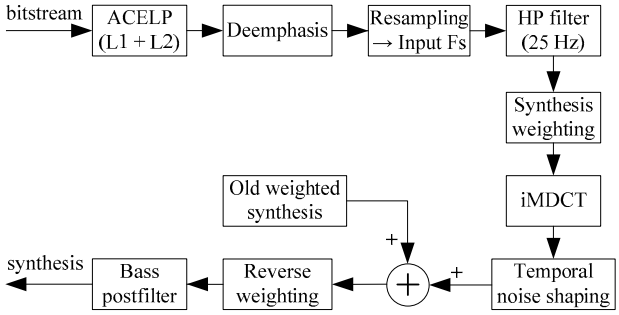

Figure 2: Structural block diagram of the decoder

4. DECODER OVERVIEW

Figure 2 shows the block diagram of the decoder. In each 20-ms frame, the decoder can receive any of the supported bit rates, from 8 kb/s up to 32 kb/s. This means that the decoder operation is conditional on the number of bits, or layers, received in each frame. In Figure 2, we assume that the output is WB and that at least Layers 1, 2, 3 have been received at the decoder.

First, the core layer and the ACELP enhancement layer (L1 and L2) are decoded. The synthesized signal is then de-emphasized, resampled to 16 kHz and high-pass filtered. Transform coding enhancement layers are added to the perceptually weighted synthesis and simple temporal noise shaping is applied. The weighted synthesis is then added to the synthesis of the previous frame with 50% overlap. Reverse perceptual weighting is applied to restore the synthesized WB signal, followed by an enhanced pitch post-filter based on [2]. The post-filter exploits the extra decoder delay introduced for the overlap-add synthesis of the MDCT layers (L3-L5).

As mentioned previously, if the decoder is limited to L2 output at call set up, a low-delay mode is used by default as the additional frame delay for MDCT overlap-add is not needed.

If L1, L2 or L3 is output by the decoder, a bandwidth extension is used to generate frequencies between 6.4 and 7 kHz. For L4 or L5 output, this is not the case anymore and the entire spectrum is quantized.

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 GC is used in the core layer is provided in Table II.

Table II. Example bit allocation for GC core layer

Layer	Param.	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 G.EV-VBR candidate codecs have been formally evaluated in ITU-T baseline selection tests. The tests showed that the most significant progress with respect to the state-of-the-art references has been made in the low bitrate WB conditions and FE conditions. Very good performance has been also achieved for NB inputs where L1 at 8 kb/s showed statistical equivalence with G.729 Annex E at 11.8 kb/s for clean speech. Finally, the codec performed very well also in noisy conditions both for NB and WB inputs. Selected results can be seen in Fig. 3 and 4 extracted from [8].

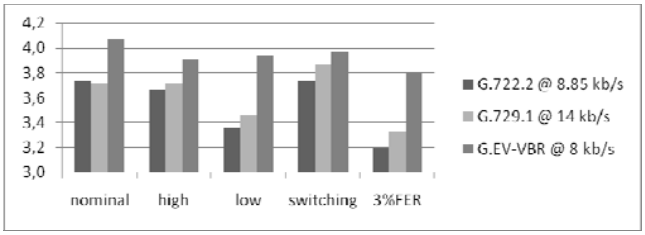

Figure 3: G.EV-VBR L1 performance comparison for WB inputs. Note : in the switching condition, G.722.2 at fixed rate of 8.85 kb/s was used, and G.729.1 rates varied between 14 kb/s and 32 kb/s only.

Fig. 3 summarizes the experiment 2a of the selection test for WB input at L1 in nominal, high and low input signal levels, 3% FE, and random switching among all the layers. The Q9 wining candidate codec is compared to the G.722.2. and G.729.1 references. It is also worth noting that the G.EV-VBR performance in 3% FE condition is close to its clean channel performance. Fig. 4 summarizes the experiment 4 and 5 where performance for NB speech at 8 kb/s and 12 kb/s was evaluated for 15 dB car noise and 20 dB street noise. The results shown are averaged from both the test and the cross-check laboratories. The comprehensive test results can be found in [8].

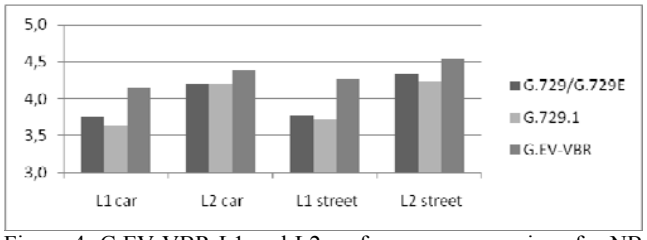

Figure 4: G.EV-VBR L1 and L2 performance comparison for NB noisy inputs. Note: G.729 Annex E operates at 11.8 kb/s.

7. CONCLUSION

We have presented the features and the architecture of the recent ITU-T Q9/16 baseline selection test winning candidate. The aim of the tests was to select a baseline codec for the development of the G.EV-VBR embedded speech and audio coding standard. Selected test results show that a major advancement has been achieved in low bit-rate WB and NB speech coding, noisy conditions, and robustness to frame erasures.

8. REFERENCES

[1] 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.

[2] M. Jelínek, 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.

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

[5] 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.

[6] 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, vol. 1, pp. 789-792, March, 2005.

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

[8] *Global Analyses for the Selection Phase for the Embedded-VBR Speech Codec*, ITU-T Q7/SG12 Technical Contribution AH-07-03 Rev.1, March 2007.