WAVELET SCALABLE SPEECH CODING USING ALGEBRAIC QUANTIZATION

M. De Meuleneire, H. Taddei

Nokia Siemens Networks
COO RTP PT HWCT Computing Techn SDE
Otto-Hahn Ring 6, 81739 Munich, Germany

D. Pastor

ENST Bretagne - Technopôle Brest-Iroise
Department SC
CS 83818, 29238 Brest Cedex 3, France

ABSTRACT

This paper proposes a new structure for a scalable codec. Our proposed codec works with 10 ms input frame for wideband speech and audio signals at bit rates ranging from 8 to 32 kbit/s. The core layer is the ITU-T G.729 at 8 kbit/s producing a narrowband output. The first enhancement layer is a bandwitdh extension providing a wideband output with 2 kbit/s. The second enhancement layer is based on algebraic quantization of wavelet packet coefficients and improves gradually the synthesized signal as the bitrate increases. For speech signals, at bitrates of 24 and 32 kbit/s, the codec is shown to be equivalent to the ITU-T G.722 codec at 56 and 64 kbit/s, respectively. Moreover, the codec at 32 kbit/s is assessed to be equivalent to the recently standardized embedded codec ITU-T G.729.1 at the same bitrate with a lower algorithmic delay.

Index Terms— embedded coding, linear predictive coding, bandwidth extension, wavelets, algebraic quantization

1. INTRODUCTION

Embedded speech coding is currently a hot topic and is being investigated in some standardization bodies as for example the ITU-T (International Telecommunication Union, Telecommunication sector). There, a scalable extension of the ITU-T G.729 [1], called G.729.1 [2] was standardized in May 2006. Besides, three other standardization activities are on going: a new embedded coder G.EV-VBR is in optimization phase, a scalable wideband extension of G.711 is also in optimization phase and a superwideband extension for both G.729.1 and G.EV-VBR is under study.

Scalable encoder generates one single bitstream and a subset can be selected at the decoder (or anywhere on the transmission path between the sender and the receiver). This has the advantage that the bitrate can be modified at anytime or any place, during the transmission, without the need for additional signaling. According to the decoding bitrate, characteristics like quality and/or bandwidth may vary.

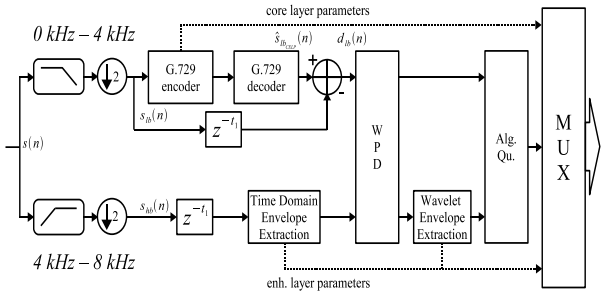
Applications such as audio conferencing may benefit from scalable coding. Participants with various terminals, e.g. personal computer, narrowband (NB) or wideband (WB) capable phones, connected through different types of networks (dial-up connection, xDSL, LAN, cable) can communicate by using a unique codec. Each terminal is able to adapt the bitrate according to its capacity and to the network traffic. Besides, to cope with network congestion, packets can be dropped on the fly, ensuring an uninterrupted conversation at the cost of quality degradation.

This paper presents a WB codec working on a 10 ms frame basis at bitrates from 8 to 32 kbit/s (maximum packet size 320 bits). This codec is a follow-up of the work proposed in [3]. The decomposition level in the Wavelet Packet Decomposition (WPD) has been decremented by one. Instead of using the Set Partitioning In Hierarchical Trees (SPIHT) algorithm, a new method to quantize the Wavelet Packet (WP) coefficients is proposed. Finally, the Bandwidth Extension (BWE) analog to the one in [2] has been replaced by a module using the WPD.

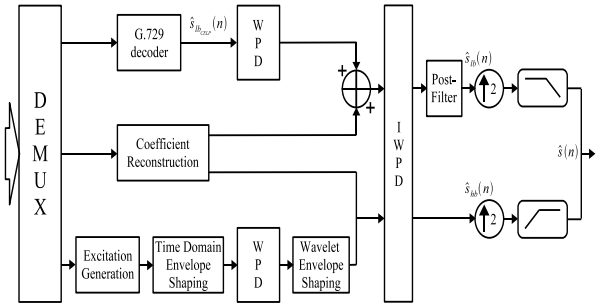
The rest of the paper is organized as follows. The encoder and decoder are detailled in Sec. 2 and 3, respectively. Results of listening tests are presented in Sec. 4. Finally, Sec. 5 concludes this paper.

2. ENCODER

The encoder is depicted in Fig. 1 (a). A WB (8 kHz bandwidth) input frame of 10 ms, i.e. 160 samples, is separated by the 2-channel QMF filter bank used in [2] into a low band (LB) part $s_{lb}(n)$, and a high band (HB) part, $s_{hb}(n)$, of 80 samples each. The LB channel is encoded by the G.729 at 8 kbit/s and locally decoded, $\hat{s}_{lb_{CELP}}(n)$. The G.729 provides the core layer parameters to the decoder. The local signal is subtracted from the original signal delayed by $t_1 = 5$ ms (introduced by the G.729 lookahead). The decoded signal $\hat{s}_{lb_{CELP}}(n)$ is subtracted from the delayed original signal $s_{lb}(n-40)$ to produce an error signal $d_{lb}(n)$. d_{lb} and s_{hb} are decomposed into WP using the 24-tap Vaidyanathan wavelet filter as in [3]. In the HB part, some BWE parameters are extracted before and after the WPD.



(a) Encoder



(b) Decoder

Fig. 1. High level schemes of the proposed codec.

2.1. BWE parameter extraction

In [3], it has been shown that a BWE tool is essential for a scalable codec in order to decode a WB signal with a constant bandwidth rendering when the decoding bitrate is varying. It avoids having holes in the spectrum when some coefficients have not been received. These holes cause artifacts similar to musical noise created by some noise reduction algorithms. Our BWE ensures a complete WB spectrum beyond 10 kbit/s.

On the encoder side, the BWE modules extracts two sets of parameters from the HB part of the original signal. These parameters correspond to the time envelope and to the frequency envelope. In order to remain time aligned with the LB part at the output of the split band structure, the HB part is delayed by 5 ms. The Time Domain Envelope Extraction module extracts the time envelope by computing a set of 8 energy values on 8 non-overlapped 1.25 ms long windows:

$$\sigma_T(j) = \sqrt{\sum_{i=0}^{9} s_{hb}^2(10j+i)}, \;\; j = 0, \ldots, 7 \qquad (1)$$

The energy within each window is normalized by the energy of the G.729 output:

$$e_{CELP} = \sqrt{\sum_{n=0}^{80} \hat{s}_{lb_{CELP}}^2(n)} \qquad (2)$$

The set of normalized energies are split vector quantized in the $\log_2$ domain with two 4-dimensional codebooks of 32 entries, at a bitrate of 10 bits per frame.

In [2] and [3], the frequency envelope extraction is done by computing a Fast Fourier Transform. In this paper, we introduce a different approach. Our frequency envelope extraction relies on the subband decomposition of the WPD. The Wavelet Envelope Extraction module extracts the frequency envelope after the WPD by computing the energy of each WP in the HB part. As the WPD comprises 4 levels, the HB part is decomposed into 8 WPs of 500 Hz bandwidth. Since the input signal is bandwidth limited (up to 7 kHz), 2 WPs are not taken into account. The energy of each remaining packet is also normalized by the energy value e_{CELP}. The vector of the normalized energies is split vector quantized in the $\log_2$ domain using two 3-dimensional codebooks of 32 entries.

Both envelopes are then applied on the decoder side to an artificial excitation signal, this is further explained in Sec. 3. The BWE parameters require 2 kbit/s (or 20 bits per frame). They constitute the first enhancement layer. 220 bits are left for the algebraic quantization of the WP coefficients.

2.2. Algebraic quantization of the wavelet coefficients

The original HB part $s_{hb}(n-40)$ and the error signal $d_{lb}(n)$ in the NB part are further encoded in the wavelet domain. The WPD provides 16 WPs. As previously explained, 2 WPs are discarded. Hence, 14 WPs only are transmitted. In [3], a very scalable quantization of the wavelet coefficients was used, namely SPIHT. Albeit highly scalable, SPIHT has not proved to be able to transmit enough coefficients to replace the coefficients from the BWE at bitrates higher than 10 kbit/s. As a result, many WP coefficients are provided by the BWE. Between 10 and 32 kbit/s the quality is not really getting better.

Instead of using SPIHT for the WP quantization, a new algebraic quantization, with the advantage of being simple, similar to the idea of algebraic codebook in CELP coding, has been developed. Let $y(i)$, $i \in \{0, \ldots, 139\}$ be the WP coefficients to be quantized. A coefficient within band k is indexed by j. A coefficient at position j in band k has position $10k+j$ in the frame. In band k, the coefficients are estimated by:

$$\tilde{y}(10k+j) = m_k c(10k+j), \;\; j \in \{0, \ldots, 9\} \qquad (3)$$

where $c(10k+j) = \pm 1$ or 0 (if $c(10k+j) = \pm 1$, it is called a pulse), and $\hat{m}_k$ is a positive value and corresponds to the pulse amplitude. At the decoder, a quantized coefficient is obtained by multiplying the amplitude m_k with $c(10k+j)$.

For a given band k, the encoder has to determine how many pulses must be sent as well as their position, their sign, and the optimal amplitude m_k. The pulse signs are given by the signs of their respective coefficients.

The pulse positions, as well as the amplitude in band k are chosen by minimizing the mean square error between the

original and the estimated coefficients:

$$e_k = \sum_{j=0}^{9} \left(y(10k+j) - m_k c(10k+j)\right)^2 \qquad (4)$$

The optimal amplitudes are given by deriving Eq. (4) according to m_k:

$$m_k = \frac{\sum_{j=0}^{9} y(10k+j)c(10k+j)}{\sum_{j=0}^{9} c^2(10k+j)} \qquad (5)$$

Since the number of pulses is a discrete number, it is not possible to compute this number by derivation (like for the amplitude). For a given number of pulses ℓ, the pulse positions to be transmitted correspond to the ℓ coefficients with the largest absolute values, as they are those that minimize the error over the possible combinations to choose ℓ pulses among 10. The error is computed for one pulse and new pulses are added as long as the error e_k decreases. When the number of optimal pulses is found, the amplitude is quantized with 4 bits using a 16-step non uniform scalar quantizer. As for the pulse positions, the encoder outputs for each WP coefficient in band k:

- 0 (1 bit) if $c(10k+j) = 0$
- 11 (2 bits) if $c(10k+j) = 1$
- 10 (2 bits) if $c(10k+j) = -1$

The WPs are transmitted according to the decreasing energy order of the 10 kbit/s output WPs (G.729+BWE). In the LB part, the energy of each band is computed on the WP coefficients of the G.729 output (core codec). In the HB part, the energy in each band is given by the BWE. The artificially bandwidth extended signal is present both at encoder and decoder side in form of WP coefficients. Let y_{BWE} be the WP coefficients of the signal decoded at 10 kbit/s. The energy of a WP k is defined by:

$$en(k) = \sum_{j=0}^{9} y_{BWE}^2 (10k+j), \qquad (6)$$

Let $k_0, \ldots, k_{13}$ be indices verifying the ordering relation:

$$en(k_0) \geq en(k_1) \geq \cdots \geq en(k_{12}) \geq en(k_{13}) \qquad (7)$$

The encoder first transmits the WP k_0 by sending the amplitude $\hat{m}_{k_0}$ and bits concerning the pulses. Afterwards, the WPs $k_1, k_2, \ldots, k_{12}, k_{13}$ are successively transmitted. For each WP, between 15 and 24 bits are needed. Consequently, a band requires in average 20 bits, that is to say the number of decoded bands increases by 1 with step of 2 kbit/s. This method has the advantage of producing a scalable codeword for each WP. Indeed, a subset of the last received codeword can be decoded to reconstruct the corresponding coefficients.

3. DECODER

The decoder is depicted in Fig. 1 (b). In the LB part, the core decoder synthesizes from the core layer parameters a 8 kHz sampling frequency signal $\hat{s}_{lb_{CELP}}(n)$ with the quality of the core codec (here G.729). Some parameters are also used by the bandwidth extension layer in the HB part. The LB part is then decomposed into WPs.

3.1. BWE decoder

The transmitted BWE parameters are used to shape a so-called excitation signal. From the gains and the codewords of the G.729 fixed and adaptive codewords, the excitation generation attempts to reproduce the fine structure of the HB part with the characteristics of the LB part, i.e. pitch values and ratio between voiced and unvoiced components:

$$exc_{wb}(n) = g_v exc_v(n) + g_{uv} exc_{uv}(n) \qquad (8)$$

with the relation $g_v^2 + g_{uv}^2 = 1$. The gain g_v is computed as in [2]. The unvoiced contribution $g_{uv} exc_{uv}(n)$ is generated by a pseudo noise generator with a gaussian distribution. As for the voiced contribution $g_v exc_v(n)$, differently from [2] and [3], it is obtained by multiplying the analytic representation of the G.729 adaptive excitation by the complex valued modulation function with the desired shift Δ_f:

$$exc_v(n) = \Re\left(\left(v(n) + \imath \mathcal{H}\left(v(n)\right)\right) e^{\imath \frac{2\pi \Delta_f n}{4000} + \Theta} \right) \qquad (9)$$

where Θ is an arbitrary phase to ensure the continuity of the modulation function and $v(n)$ is the G.729 adaptive codebook vector. To make the first regenerated harmonic and the first original harmonic coincide, Δ_f should be equal to $f_0 - (n_0 f_0 - 4000)$ Hz, f_0 is the fundamental and n_0 the first harmonic beyond 4 kHz. The analytic representation is directly related to the Hilbert transform $\mathcal{H}$ of the excitation.

The time domain envelope shaping modifies the time envelope of the excitation according to the quantized time envelope. Within a time segment as defined in Sec. 2.1, each sample is multiplied by a gain that is interpolated using a "flat window" like in [2]. This operation ensures a smooth transition over the first half of the segment as shown in Fig. 2.

In [2] and [3], the frequency envelope shaping is done by using a FIR filter bank. Our decoder uses again the WPD instead of performing the shaping. The time shaped excitation is decomposed into WPs. The energy of each WP is scaled to its respective transmitted energy (see Sec. 2.1). At this point, the WPs represent the artificially bandwidth extended signal. Should no BWE parameters be received, the output would be NB. But techniques as described in [4] could be implemented on the BWE parameters to obtain WB signal at 8 kbit/s.

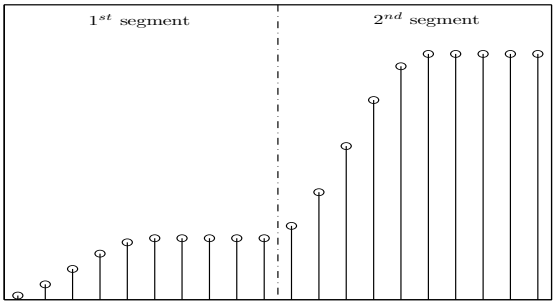


Fig. 2. Time evolution of the gain.

3.2. Reconstruction of the WP coefficients

When the bitrate is sufficiently high, the WP coefficients transmitted by the second enhancement layer are reconstructed. The higher the number of received coefficients, the higher the quality. From the 10 kbit/s WP coefficients, the decoder retrieves the order of the transmitted WPs. For each packet, the amplitude is first decoded. Afterwards, the pulses are progressively decoded. Whenever a pulse is decoded, it is multiplied by its associated amplitude. In the lower band, the decoded coefficients are added to the WP coefficients of the G.729 output. For the higher band, the missing WPs are replaced by the corresponding ones from the bandwidth extension output.

The resulting WPs in LB and HB parts are reconstructed into two channels. The low frequency channel is post-filtered using the G.729 post-filter. Finally, from the channels $\hat{s}_{lb}(n)$ and $\hat{s}_{hb}(n)$ the synthesis filter bank outputs the wideband signal $\hat{s}(n)$. The combination of the QMF and of the WPD introduces a delay of 23.125 ms. After adding the delay resulting from the G.729 lookahead (5 ms) and the one from the frame buffering (10 ms), the total arithmetic delay is 38.125 ms.

4. LISTENING TESTS

To assess the quality, listening tests for speech signals have been performed in English language. The chosen protocol is the Absolute Category Rating [5]. Eight listeners participated. None of them were native English speaker, but all speak and understand English. Six samples were presented to the listeners over headphones for two-ears listening. This is somehow more discriminative than tests conducted with one-ear listening. The proposed codec at 10, 24 and 32 kibt/s was tested against other codecs and reference signals.

We compared our codec against some references and we verified with a 95 % confidence interval whether our proposed codec was not worse than (n.w.t) the reference. Results are gathered in Tab. 1. Diff. and Thr. stands for the MOS difference and the comparison threshold, respectively. The column Result shows whether the requirement is passed (Diff.$\leq$Thr.).

The proposed codec at 10 kbit/s (Codec A1) was tested against the AMR-WB at 12.65 kbit/s (Codec B). At this bitrate, we do not reach the quality of the AMR-WB. This was expected as we have less bitrate. In addition, our structure is scalable. It would have been better to compare with AMR-WB at 8.85 kbit/s, but we wanted to keep a reasonable test size. At 24 kbit/s (Codec A2) and 32 kbit/s (Codec A3), our codec is proved to be at least equivalent to the G.722 at 56 kbit/s (Codec C1) and 64 kbit/s (Codec C2) respectively. Although the G.729.1 at 24 kbit/s (Codec D1) seems to be better than the proposed coder at the same bitrate, the gap is filled in at 32 kbit/s (Codec D2).

Requirement	Y_{test}	Y_{ref}	Diff.	Thr.	Result
A1 n.w.t B	3.21	3.90	0.69	0.16	No
A2 n.w.t C1	3.94	4.10	0.16	0.20	Yes
A2 n.w.t D1	3.94	4.33	0.39	0.17	No
A3 n.w.t C2	4.13	3.94	-0.19	0.19	Yes
A3 n.w.t D2	4.13	4.27	0.14	0.20	Yes

Table 1. Tested conditions on speech material.

5. CONCLUSION

This paper discussed a 8-32 kbit/s scalable WB speech coder. The 10 ms WB input frame is split into two components. The LB part is encoded with the G.729 that produces the core layer. The HB part is analyzed to extract the parameters that constitute the first enhancement layer necessary to the BWE. The error between the delayed original signal and the G.729 output in the LB part, and the delayed original in the HB part are decomposed into WPs. The quantized coefficients provide the second and last enhancement layer that is also scalable. The listening test has shown that the quality of the codec improves gradually as the bitrate increases. In clean speech conditions, the quality is close to that of the G.729.1 (20 ms frame length, 48.9375 ms algorithmic delay), that is a state of the art codec. However, it was achieved with a shorter frame length (10 ms) and a lower algorithmic delay (38.125 ms).

6. REFERENCES

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

[2] ITU-T, "Recommendation G.729.1: G.729 Based Embedded Variable Bit-Rate Coder: An 8-32 kbit/s Scalable Wideband Coder Bitstream Interoperable with G.729," May 2006.

[3] M. De Meuleneire, H. Taddei, O. de Zelicourt, D. Pastor, and P. Jax, "A CELP-Wavelet Scalable Wideband Speech Coder," in *Proc. of ICASSP 2006*, Toulouse, France, May 2006, IEEE, vol. 1, pp. 697–700.

[4] B. Geiser, H. Taddei, and P. Vary, "Artificial Bandwidth Extension without Side Information for ITU-T G.729.1," in *Proc. of Interspeech 2007*, Aug. 2007.

[5] ITU-T, "Recommendation P.800: Methods for Subjective Determination of Transmission Quality," Aug. 1996.