A CELP-WAVELET SCALABLE WIDEBAND SPEECH CODER

Mickaël De Meuleneire, Hervé Taddei†, Olivier de Zélicourt†, Dominique Pastor‡, Peter Jax⸝

BenQ Mobile, Munich, Germany (mickael.demeuleneire@benq.com)
†Siemens AG, Munich, Germany (firstname.name@siemens.com)
‡ENST Bretagne, Brest, France (dominique.pastor@enst-bretagne.fr)
⸝IND, RWTH Aachen University, Germany (jax@ind.rwth-aachen.de)

ABSTRACT

This paper presents a scalable wideband speech codec working at bitrates ranging from 8 to 32 kbit/s. The core layer is the ITU-T G.729 at 8 kbit/s. A first enhancement layer is a bandwidth extension algorithm requiring 2 kbit/s to widen the G.729 narrow band output speech. The difference between the wideband original and reconstructed signal is transformed in the time-frequency domain by a full wavelet decomposition. The resulting coefficients are quantized by an embedded quantizer at 22 kbit/s. Listening tests show the relevance of such a scheme when compared to a pure wavelet packet decomposition. In addition, listening tests suggest that the proposed codec is equivalent to the ITU-T G.722 at 48 kbit/s for speech signals.

1. INTRODUCTION

Scalable coding refers to methods where the output of the encoder, that is to say the bitstream, is organized in layers. The *core layer* comprises the parameters needed to reconstruct audio signals with minimal quality at the lowest bitrate. If this layer is not received, then the signal cannot be reconstructed and frame erasure concealment should be applied. The other layers, called *enhancement layers*, include parameters to improve the reconstructed signal either by working on the available decoded signal or by increasing the signal bandwidth (e.g. from narrowband to wideband). Layers are then transmitted according to the available bandwidth. On the decoder side, every decoded layer increases the quality of the reconstructed signal.

Scalable coding is particularly suitable for delivering contents to different types of networks (e.g. dial-up connection, xDSL, LAN) and hardware (e.g. PC, narrowband or wideband capable phones) and especially Packet Switched networks and Voice over IP applications. To reduce network congestion or to increase the number of users over a backbone, some entities in the network may discard the higher layers. Unequal error protection can very easily be implemented with a simple scheme where, for example, the core layer is better protected than the other layers.

Many types of scalable codecs exist. LPC-based (Linear Predictive Coding) scalable codecs focus on quantizing a residual excitation [1]. Usually the residual excitation at the lowest bitrate is refined by adding the contribution of further codebooks to reduce artefacts or distortion caused by lower enhancement layers. The quality does not increase linearly with the number of codebooks but reaches a saturation point after a few codebook contributions [2].

Hybrid scalable codecs usually comprise a narrowband or wideband CELP (Code Excited Linear Prediction) as a core codec [3]. The enhancement layers encode the error between the original and the reconstructed signal in the frequency domain. By taking into account the masking properties of the human ear, psychoacoustic models can be introduced to better quantify the perceptually most important coefficients by allocating more bits to them.

The structure we propose derives from this approach. We combine CELP coding and wavelet transform. Wavelets by themselves have been applied succesfully to speech and audio coding, with or without scalable structures [4][5][6]. In the following, we first describe the structure of the proposed coder. After a short presentation on wavelets used as filter banks for scalable coding, we present some performance and quality evaluations. In particular, we discuss the relevance of using a speech model for the core layer. Finally, we mention some potential future work and draw conclusions.

2. OUR CELP-WAVELET CODER SCHEME

The encoder scheme is depicted in Fig. 1. The coder works with a 10 ms frame length and 16 kHz input signal. First, the input signal is downsampled to 8 kHz. This signal is then used as the input of the core layer, in our case the ITU-T G.729 [7] at 8 kbit/s. It provides the core layer parameters. These parameters are locally decoded to obtain a narrowband decoded signal. A Time Domain Band Width Extension (TDBWE) algorithm [8] widens the narrowband output signal of the G.729 to a wideband signal. This was proposed by Matsushita, Mindspeed and Siemens in their ITU-T G.729EV candidate at 14 kbit/s mode and uses 2 kbit/s. The difference

between the TDBWE output and the delayed original signal is transformed by a wavelet packet decomposition. Finally, the coefficients are encoded by an embedded quantizer, which enables their progressive transmission as explained in section 3.2. The maximal bitrate is fixed to 32 kbit/s. Thus 22 kbit/s are available for the quantization of the coefficients.

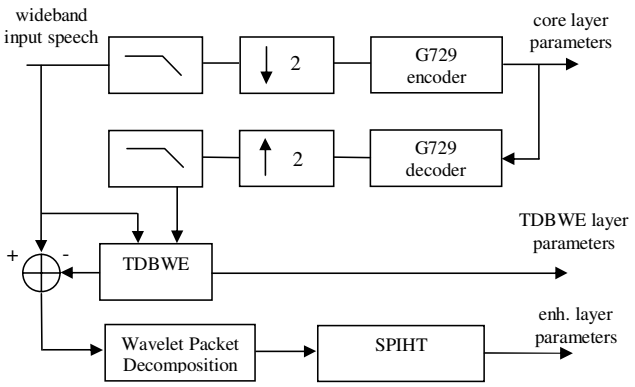

Fig. 1. Structure of the proposed coder.

We used 24-tap Vaidyanathan wavelet filters [9] as these filters provide a good trade-off between filter length and frequency selectivity. We chose to perform a full wavelet packet decomposition on 5 levels. We then obtained 32 packets of 5 samples which introduces a delay of 682 samples (or 43.6 ms) as explained in 3.3. Additional delays include the algorithmic delay of the two low pass filters (41-tap filter) of 2.5 ms, 5 ms for the look ahead of the core layer, 1.25 ms for the TDBWE part and 10 ms for the frame buffering. This leads to a total arithmetic delay of 62.35 ms.

3. WAVELETS FOR SCALABLE CODING

In this section, we start by giving a short description of wavelets used as filter banks, as well as examples of embedded quantizer for wavelet coefficients. We detail the structure and the behaviour of our codec with respect to standard issues in speech/audio coding.

3.1. Wavelet transform

The wavelet transform belongs to the family of filter banks. It consists of a low-pass filter and a high-pass filter followed by a decimation of factor 2 (Fig. 2). For an N-sample input frame, the two $N/2$-sample output frames are called approximation and detail, with a frequency bandwidth roughly half that of the input signal. The low frequency band (the approximation) can be further halved, providing details at lower resolution levels. The details are not further filtered. The synthesis filters derived from the analysis filters ensure perfect reconstruction. This method favours good frequency selectivity for low frequencies at the cost of the temporal resolution. The opposite is observed for the high frequency part. This poor frequency selectivity in the high frequency area is not really acceptable when doing speech/audio coding. That is why our codec is based on wavelet packet (WP) decomposition.

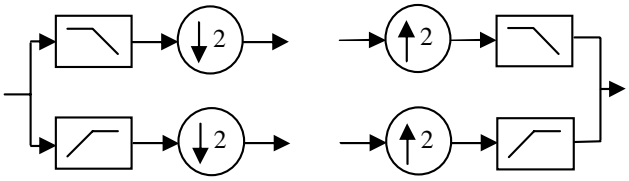

Fig. 2. Analysis/synthesis filter bank.

The WP decomposition is an extension of the wavelet transform. With this method, details are further decomposed into approximations and details (approximation of details and details of details, cf fig. 3). This amounts to halving the upper frequency bands. Since the G.729 and TDBWE provide a good temporal resolution, we perform a full WP decomposition on the difference to obtain the best possible frequency selectivity. The wavelet decomposition is usually performed on dyadic segments, i.e. the segment lengths are a power of two. The decomposition is then performed as far as possible, that is to say until only one coefficient is left available in each packet, to get the best possible frequency selectivity. However, there is no limitation on performing the decomposition on non-dyadic frames. The decomposition is stopped when the segment length is odd. If the length ℓ of the segment is $\ell = k.2^M$ with k odd, the decomposition may be performed M times. In our case, as we have 160 samples, we get $k = 5$ and $M = 5$. More details about wavelets can be found in [10].

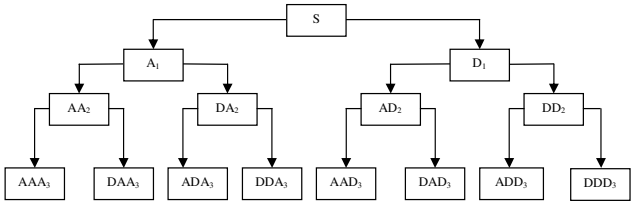

Fig. 3. Wavelet packet decomposition.

3.2. Scalability

Scalable coding of wavelet coefficients can be achieved by using either the Embedded Zero-tree Wavelet (EZW) [11] or the Set Partioning In Hierarchical Tree (SPIHT) [12] algorithm. Both have been applied successfully for still image compression based on wavelet transforms. A description of these algorithms for an 1-D signal can be found in [13] and [14] respectively. In audio coding, we exploit the assumptions that most of the signal energy is concentrated in the low

frequency bands and that the wavelet coefficients are correlated across subbands. This second assumption is often called "self-similarity". The coefficients are organized as a tree from coarse to fine scale (i.e. from low to high frequencies).

EZW and SPIHT progressively quantize the coefficients from the most significant to the less significant bits. Coefficients with the largest amplitudes are quantized first, the smallest ones last. Coefficients are classified according to their significance. A coefficient c_n is said to be significant, if for a given integer k: $|c_n| \geq 2^k$, otherwise it is insignificant (2^k is the so called threshold). To some extent, a subset of coefficients is called significant if at least one of these coefficients is significant, otherwise it is insignificant (also called a zero-tree).

Although EZW and SPIHT are based on the same assumptions, they differ in the way they scan and partition the coefficients. Partitioning rules and scanning orders are shared by the encoder and the decoder. The encoding as well as the decoding process can be stopped at any time, e.g. when a target bitrate is reached. A few decoded bits give a coarse representation of the coefficients. The more bits that are decoded, the more refined the representation is. SPIHT is known to yield better performances.

Table 1 (see next page) gives some segmental Signal to Noise Ratio (SNR) results. The segmental SNR is a measure of the distortion between the original and the synthesized signal. We compared EZW with SPIHT for different schemes at 24 and 32 kbit/s. In the first scheme (scheme A), the wavelet decomposition is applied directly to the original signal. In the second scheme (scheme B), the wavelet decomposition is applied to the difference between the original signal and the output of the G.729. The third scheme (scheme C, see section 2) consists of the wavelet decomposition applied to the difference between the original signal and the output of the TDBWE. We note that the distortion decreases as the bitrate increases. Table 1 confirms that SPIHT is generally better than EZW and especially when the bitrate increases. Nevertheless, informal listening tests did not show a significant quality difference between SPIHT and EZW.

3.3. Reconstruction problems

The simplest implementation of wavelet transforms and packet decomposition is circular convolution [10]. Since the convolution needs information outside the segment to be filtered, the data are expanded periodically. The method guarantees perfect reconstruction of the output samples, but creates larger coefficients on the edges and can cause undesired edge effects when wavelet coefficients are coarsely quantized. One alternative is to consider for the current frame the input samples and the coefficients from the past frames [5]. This method was called "full convolution" in [15]. For an N-tap filter and an L-level WP decomposition, the delay introduced by this convolution is $(2^L - 1)(N - 2)$.

The WP decomposition can be applied to a difference between the original wideband signal and the narrowband reconstructed signal. This was the strategy of most proponents during the ITU-T G.729EV competition. Consequently, the coefficients corresponding to the high frequency part represent the original signal. Depending on the available bitrate for these frequency bands, the coefficients can be quantized or not depending on the constant global bitrate. When these coefficients are not available at the decoder side, this leads to artefacts similar to the musical noise created by some noise reduction algorithms. To cope with this effect, we use the TDBWE algorithm. The input of the WP decomposition is then a difference between two wideband signals. When wavelet coefficients are missing at the decoder side, the corresponding spectral part is provided by the TDBWE layer.

4. RESULTS OF LISTENING TESTS

In this section we evaluate the impact of using a speech model core codec, that is to say the G.729, together with TDBWE. Our goal is to measure how these modules contribute to the wavelet decomposition and reconstruction. For this purpose, different A-B listening tests were performed: 9 people listened to 12 pairs of samples, 8 speech and 4 music. The speech material is in French and comes from the NTT database. The bitrate was fixed to 32 kbit/s. First, we wanted to compare the schemes A and B. This pure wavelet codec A consists of a 7 level WP decomposition on a 256 sample frame basis encoded by SPIHT. We use 256 samples as the quality is better than when using 160 input samples. This slightly favoured the pure wavelet codec. The results are presented in table 2. They represent the preference for one of the coders over the other one. The results clearly show that using

Signal	A:WP	B:CELP+WP
Speech	7 %	93 %
Music	11 %	89 %

Table 2. Improvement thanks to CELP coding in the coder.

Signal	B:CELP+WP	C:CELP+TDBWE+WP
Speech	19 %	81 %
Music	25 %	75 %

Table 3. Intregration of the TDBWE.

a narrowband codec G.729 as the core codec yields a better quality at 32 kbit/s than when using a pure wavelet codec in about 90% of the cases. Then we checked whether it was better to add TDBWE to the G.729 codec and then to perform wavelet coding (scheme C). Finally, the use of TDBWE increases the quality of the reconstructed signal for both speech and music in about 80% of the cases (see Tab. 3). Since the G.729 is dedicated to speech coding, it is more efficient than

Signal	A:WP		B:G.729+WP		C:CELP+TDBWE+WP	
	EZW	SPIHT	EZW	SPIHT	EZW	SPIHT
music@24k	11,91 dB	14,57 dB	12,51 dB	13,95 dB	11,88 dB	12,81 dB
music@32k	14,41 dB	17,82 dB	14,73 dB	16,87 dB	14,21 dB	15,69 dB
male@24k	18,83 dB	19,31 dB	16,85 dB	18,28 dB	16,34 dB	16,09 dB
male@32k	21,31 dB	21,97 dB	18,92 dB	20,91 dB	18,35 dB	18,46 dB
female@24k	18,86 dB	19,68 dB	17,72 dB	19,29 dB	17,24 dB	17,13 dB
female@32k	21,09 dB	22,20 dB	19,42 dB	21,68 dB	19,00 dB	19,22 dB

Table 1. Comparison based on segmental SNR improvement.

the wavelet coder at the same bitrate. As for TDBWE, it provides the spectrum parts that are missing when these wavelet coefficients are not decoded or transmitted due to the bit allocation of the SPIHT algorithm. TDBWE ensures a constant frequency bandwidth of the reconstructed signal. Then we took the decision to include TDBWE together with the G.729 in our scalable codec scheme. We conducted further informal listening tests to compare our coder at 32 kbit/s with the ITU-T G.722 at 48 kbit/s and 56 kbit/s in the same conditions as the previous test. The results are presented in table 4. Results

Signal	G.722 48 kbit/s	Scheme C	G.722 56 kbit/s	Scheme C
Speech	50 %	50 %	59.8 %	40.2 %
Music	77.8 %	22.2 %	77.8 %	22.2 %

Table 4. Comparison with G.722 at 48 kbit/s and 56 kbit/s.

for speech are promising, since the coder is equivalent to the G.722 at 48 kbit/s, and close to the G.722 at 56 kbit/s. But still further work is needed to increase the quality. For music, it can be seen that the quality is very reduced due to the speech model used by the core codec. The quality for music is probably impacted by the size of the frames (10 ms). Usually audio codecs use longer frame sizes. We are currently evaluating ways to improve the quality, like for example arithmetic coding of the SPIHT output or use of a psychoacoustic model.

5. CONCLUSION

We have proposed a wideband scalable coder using the ITU-T G.729 as the core layer at 8 kbit/s. A first enhancement layer provides the upper frequency band of the wideband signal using 2 kbit/s. The final enhancement layer uses a WP decomposition together with an embedded quantizer which enables progressive decoding between 10 and 32 kbit/s. For the targeted bitrates we show the improvements due to the integration of CELP coding and bandwidth extension in a wavelet based scalable coder. Informal listening tests show that the coder at 32 kbit/s for speech is equivalent to the ITU-T G.722 at 48 kbit/s. Further optimizations are needed to meet the objective of being at least equivalent to the G.722 at 56 kbit/s.

6. REFERENCES

[1] A. Le Guyader, C. Lamblin, and E. Boursicaut, "Embedded Algebraic CELP/VSELP Coders for Wideband Speech Coding," *Speech Communication*, vol. 16, no. 4, pp. 319–28, 1995.

[2] C. Erdmann and P. Vary, "Performance of Multistage Vector Quantization in Hierarchical Coding," *European Transactions on Telecommunications*, vol. 4, no. 15, pp. 363–372, July/August 2004.

[3] B. Kövesi, D. Massaloux, and A. Sollaud, "A Scalable Speech and Audio Coding Scheme With Continuous Bitrate Flexibility," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 1, pp. 273–276, May 2004.

[4] D. Sinha and A. H. Tewfik, "Low Bitrate Transparent Audio Compressing Using Adapted Wavelets," *IEEE Trans. Signal Proc.*, vol. 41, pp. 3463–3479, December 1993.

[5] Y. Karelic and D. Malah, "Compression of High-Quality Audio Signals Using Adaptive Filter Banks and a Zero-Tree Coder," in *Electrical and Electronics Engineers in Israel*. IEEE, March 1995.

[6] G. Dongme, H. Wen and W. Jiangqin, "Complexity Scalable Audio Coding Algorithm Based on Wavelet Packet Decomposition," *ICSP 2000*, vol. 2, pp. 659–665, August 2000.

[7] International Telecommunications Union, "Coding of Speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)," March 1996.

[8] P. Jax, B. Geiser, S. Schandl, H. Taddei, and P. Vary, "An Embedded Scalable Wideband Codec based on the GSM EFR Codec," *to be published in ICASSP2006*, May 2006.

[9] P.P. Vaidyanathan and P.Q. Hoang, "Lattice Structures for Optimal Design and Robust Implementation of Two-Channel Perfect Reconstruction Filter Banks," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 36, pp. 81–94, January 1988.

[10] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, second edition, 1998.

[11] J. M. Shapiro, "Embedded Image Coding Using Zero-tree of Wavelet Coefficients," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, December 1993.

[12] A. Said and W. A. Pearlman, "A New Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, pp. 243–250, June 1996.

[13] Y.S. Mo, W. S. Lu, and A. Antoniou, "Embedded Coding for 1-D Signals Using Zero-trees of Wavelet Coefficients," *IEEE Pacific Rim Conf. Comm., Comp. and Signal Processing*, pp. 306–309, August 1997.

[14] Z. Lu, D. Y. Kim, and W. A. Pearlman, "Wavelet Compression of ECG Signals by the Set Partitioning in Hierarchical Trees (SPIHT) Algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 47, pp. 849–856, July 2000.

[15] B. Leslie and M. Sandler, "A Wavelet Packet Algorithm for 1-D Data with no Block End Effects," *IEEE International Symposium on Circuits and Systems VLSI*, vol. 3, pp. 423–426, June 1999.