REAL TIME IMPLEMENTATION OF THE HVXC MPEG-4 SPEECH CODER

Alessandro Maso *Cosmo Trestino* *Gian Antonio Mian*

Dept. of Electronics and Informatics
University of Padova
sashia@dei.unipd.it wafer@dei.unipd.it mian@dei.unipd.it

ABSTRACT

In this paper we present the results of the code optimization for the HVXC MPEG-4 speech coder. Two kinds of bit-rate formats are considered: 2 and 4 kbit/s. After a short description of the HVXC main features, results of code optimization are reported: the real time implementationon, on a floating point DSP, of three parallel 2 kbit/s or two parallel 4 kbit/s HVXC coders, is shown to be possible.

1. INTRODUCTION

Reduction of bit-rate in speech applications, can imply poor quality. Typical coding as Code-Excited Linear Predictive coding (CELP, also called Vector Excitation Coding VXC) [1] presents low intelligibility below 4 kbit/s. One major cause of CELP quality degradation, is due to inefficient harmonic representation. To improve quality of these vocoders, hybrid coding can be used. Harmonic and Vector eXcitation Coding (HVXC) [2] uses principles of CELP coders [1] and Multi-Band Excitation (MBE) coders [3]. The HVXC coder classifies the input signal into four classes: a) unvoiced; b) mixed with low level voicing; c) mixed with high level voicing; d) voiced. If the input signal is classified as unvoiced, a closed loop search as in vector excitation coding is carried out. If the signal is classified with some voicing degree, vector quantization of the spectral envelope is used.
At the decoder, an unvoiced component synthesizer is used for unvoiced signal, while for a signal with some voicing degree a mixed noise plus voiced component synthesizer is used. In this way a fine representation of voiced-unvoiced and unvoiced-voiced transitions is obtained.
The HVXC MPEG-4 vocoder results in high quality speech (sampled at 8 kHz) at very low bit-rate: the 2 kbit/s HVXC coder quality is comparable with that of the 4.8 kbit/s FS1016 coder [4].

2. HVXC CODER

The HVXC coding scheme is depicted in Fig. 1. The Linear Predictive Coding (LPC) block extracts the parameters of the autoregressive model. LPC parameters are converted to LSP parameters that are vector quantized with a partial prediction and multistage vector quantization scheme [5].
The open-loop pitch block estimates pitch frequency while the spectral envelop block refines pitch and extracts the harmonic amplitudes. The closed-loop section is a regular CELP [1]. The voiced-unvoiced block classifies the current 20 ms long frame into one of the four previously mentioned possible classes.

The use of both time and frequency domains for speech signal analysis, in particular the necessity to switch from one domain to the other, and the search into codebooks are the main causes of the HVXC coder computational complexity.

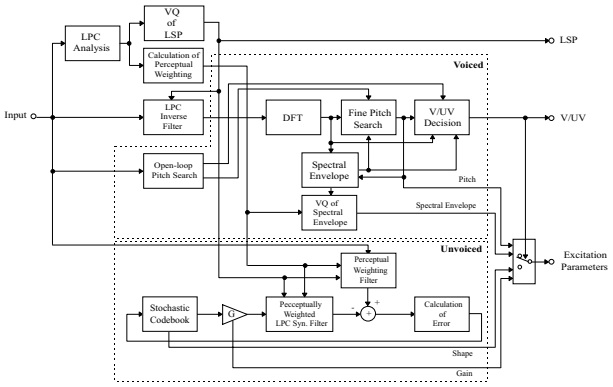

Figure 1: *Block diagram of HVXC coder.*

2.1. Pitch analysis

The open loop pitch value is estimated from the peak values of the LPC residual autocorrelation; to have a continuous pitch contour past and current pitch values are considered.

The block diagram of the pitch extractor is depicted in Fig. 2, and it can be divided into three main section :

1- extraction of energy balance and signal level[6] necessary to subsequent discrimination;

2- two parallel paths for the pitch estimation from the LPC residual autocorrelation; the difference between the two paths is the high-pass or low-pass residual filtering. Each path gives a parameter set useful for pitch estimation and also for V/UV decision;

3- discrimination between the two pitch estimates.

2.2. Harmonic analysis

In the HVXC coder, the quality of speech is mainly due to the estimation of the amplitude of the harmonics. The approach derives from the MBE vocoder [3]. The magnitude estimation is carried out by computing an optimal amplitude A_m ot the m^{th} harmonic from the Fourier transform of the LPC residual $S(\omega)$ using a predefined window. The amplitude estimation error E_m of the m^{th} harmonic is given by :

$$E_m = \frac{1}{2\pi} \int_{a_m}^{b_m} [|S(\omega)| - |A_m H(\omega)|]^2 d\omega \qquad (1)$$

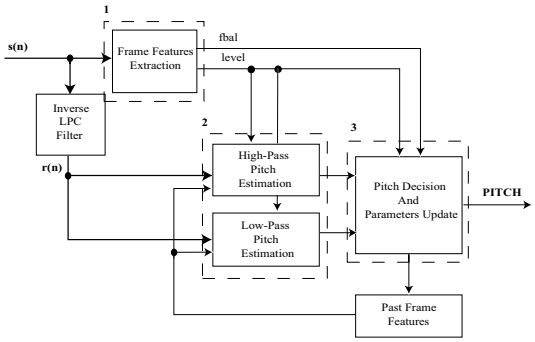

Figure 2: *Block diagram of pitch extractor.*

where a_m and b_m delimit the interval centered on the m^{th} harmonic and $H(\omega)$ denotes the Fourier transform of the Hamming window. Minimization of the amplitude estimation error, yields the harmonic amplitude A_m:

$$|A_m| = \frac{\int_{a_m}^{b_m} |S(\omega)||H(\omega)|d\omega}{\int_{a_m}^{b_m} |H(\omega)|^2 d\omega} \tag{2}$$

It can be noted that, in a different way respect to the original MBE [3], the HVXC coder uses the Fourier transform of a Hamming window, instead of the excitation spectrum.

The harmonic spectral envelope for voiced frames is then vector quantized. The variable dimension spectral vector is first converted into a fixed dimension vector by bandlimited interpolation [6] and then quantized using a weighted distortion measure [7]

2.3. Voiced-unvoiced decision

To classify the voicing degree a trained neural network is used. Signal power, harmonic structure of power spectrum, maximum autocorrelation of LPC residual and number of zero crossing are the discriminant parameters; moreover the modulo of residual in the frequency domain, is used to develop a simplified V/UV model [6]. In the formulation of the principles of the MBE principles [3], Griffin D. W. and Lim J. S. present the excitation spectrum as a periodic spectrum that can be interlaced with a noise spectrum; the simplified model implies that just one transition point between periodic spectrum and noise spectrum is allowed.

2.4. CELP analysis

If the signal segment is unvoiced, regular CELP coding, using stochastic codebook, is carried out. If c_k indicates the excitation vector at index k, the mean square weighted error is given by:

$$E_k = \|\mathbf{x} - g\mathbf{H}c_k\|^2 \tag{3}$$

where $\mathbf{x}$ is a target vector given by the weighted input speech after subtracting the zero-input response of the weighted synthesis filter, g is the gain factor and $\mathbf{H}$ is a lower triangular convolution matrix constructed from the impulse response of the weighted filter. Minimization of (2) yields:

$$g = \frac{\mathbf{x}^T\mathbf{H}c_k}{c_k^T\mathbf{H}^T\mathbf{H}c_k} \tag{4}$$

and the optimum codeword is selected maximizing the term:

$$\frac{(\mathbf{x}^T\mathbf{H}c_k)^2}{c_k^T\mathbf{H}^T\mathbf{H}c_k} \tag{5}$$

3. HVXC DECODER

The HVXC decoder scheme is depicted in Fig. 3. The decoding process is based on three steps: parameters de-quantization, voiced-unvoiced excitation signal synthesis and LPC synthesis.

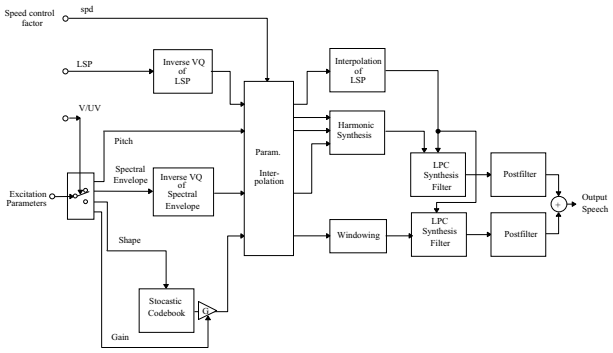

Figure 3: *Block diagram of HVXC decoder.*

3.1. Voiced component synthesizer

The construction of the voiced excitation consists of two steps: harmonic excitation synthesis and noise component addition as reported in Fig. 4; in this way a noise component is added to a periodic waveform and the resulting excitation signal is sent into the LPC synthesis filter. Great part of the quality in the reconstructed speech signal, is due to the synthesis of the voiced excitation signal. As it can be noted in Fig. 4 harmonic synthesis via IFFT is done [7]; this permits to reduce the synthesizer complexity respect to the original formulation of the MBE vocoder [3], nevertheless a remarkable complexity remains.

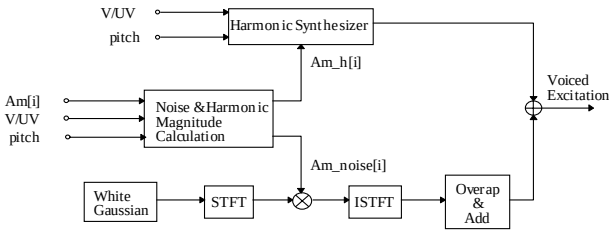

Figure 4: *Block diagram of voiced component synthesizer.*

3.2. Unvoiced component synthesizer

The unvoiced excitation is given by the stochastic vector representation of the residual and it is sent into the LPC synthesis filter. The complexity of the unvoiced component synthesizer can be neglected respect to the one of the voiced component synthesizer.

4. OPTIMIZATION

In the HVXC coder the evaluation of spectrum properties, as harmonic amplitude or energy, is of fundamental importance; voiced/unvoiced discrimination for the input speech signal, pitch refinement, harmonic amplitude extraction, are examples of parameters extracted from input speech signal and evaluated in frequency domain. Autocorrelation and number of zero crossings are examples of parameters evaluated in time domain. Because the Fourier transform is used in different parts of the HVXC coder, the careful implementation of a Fast Fourier Transform for real data, greatly increases the coder performance, as reported in table 1:

Function	Clock cycles
Original Fourier transforms	400.000
Optimized Fourier transforms	160.000
Performance increment	$\frac{240.000}{400.000} = 60\%$

Table 1: *FFT optimizations*

In the HVXC coder, the principles of the MBE coder are the bases for the discrimination of the speech signal spectrum in voiced/unvoiced (or periodic/noise) and the estimation of harmonics amplitude. Referring to this last point, the difference between HVXC and MBE coders is in the function used as excitation spectrum. While in the MBE coder the periodic excitation spectrum is derived from the signal and changes depending from the signal, in HVXC coder to represent periodic excitation spectrum a predefined window function, as Hamming window, is used.
A numerical computation of the integral in (2) gives:

$$|A_m| = \frac{\sum_{n=N_{a_m}}^{N_{b_m}} |S(n)||H(n)|}{\sum_{n=N_{a_m}}^{N_{b_m}} |H(n)|^2} \tag{6}$$

where N_{a_m} and N_{b_m} stay for the samples associated to a_m and b_m, respectively.

Because the Fourier transform of the Hamming window is a priori known, its values can be stored. Moreover, because the pitch values are limited ($20 \div 147$ corresponding to $50 \div 400$ Hz) it is possible to store the values of the denominator inverse and to avoid divisions.
This permits to obtain the performance increment shown in table 2:

Function	Clock cycles
Original harmonic extractor	850.000
Optimized harmonic extractor	130.000
Performance increment	$\frac{720.000}{850.000} \approx 85\%$

Table 2: *Harmonic extractor optimizations*

In the harmonic quantization, the search in codebook implies the maximization of a term with the presence of a square root term at the denominator:

$$max_s \ \frac{\sum_n w(n)^2 x(n) s(n)}{\sum_n \sqrt{w(n)^2 s(n)^2}} \tag{7}$$

where $w(n)$ is a perceptual weight function, $x(n)$ represents the harmonics after dimension conversion, and $s(n)$ represents the harmonics in the codebook. Because $s(n)$ is the sum of two terms (two codebook words), $s_1(n)$ and $s_2(n)$, a three loop cycle is necessary to find the optimum value for s. In the HVXC harmonic quantization two loops of 16 cycles and one loop of 44 cycles are presents; so if only one operation is present in the internal loop $16 \times 16 \times 44 = 11,264$ operations are required. The removal of the square root computation in the internal loop implies an increment in performance. This removal and the code re-arrangement in Dimension Conversion (D.C.) and Harmonic Vector Quantization (H.V.Q.), permit to reduce execution time as reported in table 3:

Function	Clock cycles
Original D.C and H.V.Q	750.000
Optimized D.C. and H.V.Q	160.000
Performance increment	$\frac{590.000}{750.000} \approx 79\%$

Table 3: *D.C. and H.V.Q. optimizations*

In addition, where the evaluation of the magnitude of vector quantities is needed, an approximation method was used [8]. Typical magnitude evaluation of $R + jI$, implies the computation of the square root of the squares of R and I :

$$M = \sqrt{R^2 + I^2} \tag{8}$$

The objective of the approximation is to determine the maximum and the minimum between $|R|$ and $|S|$, and to find two constants, a and b, such that the following approximation holds:

$$M \approx \widetilde{M} = a(max(|R|, |S|)) + b(min(|R|, |S|)) \tag{9}$$

We have chosen two sets of values for a and b :

- $a = 0.996, b = 0.123$ if $max(|R|, |S|) \geq 4min(|R|, |S|)$
- $a = 0.886, b = 0.502$ if $max(|R|, |S|) < 4min(|R|, |S|)$

The use of the magnitude approximation technique permits a (small) performance increment as reported in table 4:

Function	Clock cycles
HVXC	990.000
HVXC (magnitude approximation)	950.000
Performance increment	$\frac{40.000}{990.000} \approx 4\%$

Table 4: *magnitude approximation optimization*

The error between original and approximated magnitude was less than 2% for all frames of a 10 minutes speech database.

Similar optimizations were applied to the decoder, and gave rise to a 58% improvements. However ad hoc optimizations of the "Weighted Overlap and Add" (W.O.A.) operation of figure 4 is done and the result is a 91% clock cycles reduction, as shown in table 5:

Function	Clock cycles
Original W.O.A	350.000
Optimized W.O.A.	30.000
Performance increment	$\frac{320.000}{350.000} \approx 91\%$

Table 5: *Weighted overlap and add optimization*

The global performance improvements are reported in table 6 for both coder and decoder respectively :

Function	Clock cycles
Original HVXC coder	6.000.000
Optimized HVXC coder	950.000
Performance increment	$\frac{5.050.000}{6.000.000} \approx 84\%$
Original HVXC decoder	2.300.000
Optimized HVXC decoder	970.000
Performance increment	$\frac{1.330.000}{2.300.000} \approx 58\%$

Table 6: *HVXC vo-coder optimization*

5. RESULTS

The implementation of the HVXC vocoder has been carried out on the TMS320C6711 floating point DSP.
For a 20 ms frame, real time execution on such a DSP with a 150 MHz clock implies that the global number of clock cycles be less than 3×10^6 for each frame. At the coder side the search on three codebooks for vector quantization in LPC, CELP and Harmonic analysis is a very complex task; moreover pitch estimation and harmonic amplitudes estimation, represent the two other tasks with high computational complexity. At the decoder side the voiced component synthesis represents the more heavy task. This is clearly shown by the bar plots on the left of Fig. 5 and 6 that give, for the coder and decoder, respectively, the per-frame clock cycles needed by the C language reference code downloadable at the MPEG AUDIO official site [9].
Careful rewriting of the main functions, gave rise to the results shown in the bar plots on the right of the figures, e.g., the coder complexity is reduced to one sixth. Notice that such results were obtained without resorting to assembly language and to fixed point arithmetic. Actually the new C language version produces both the same bit-stream and the same synthesized signal as the reference software.
For the 2 kbit/s HVXC coder, actual execution time is 0.95×10^6 per-frame clock cycles, while for the 4 kbit/s HVXC coder it is 1.35×10^6; the difference is due to different search in codebooks. At the decoder side, execution time is 0.97×10^6 clock cycles for both 2 kbit/s and 4 kbit/s bit-rates. A real time implementation of three parallel 2 kbit/s or two 4 kbit/s HVXC coders is possible.

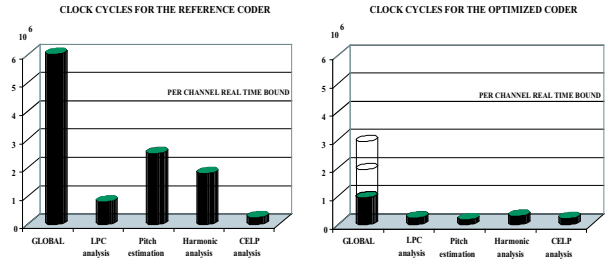

Figure 5: *Per-frame clock cycles of coder.*

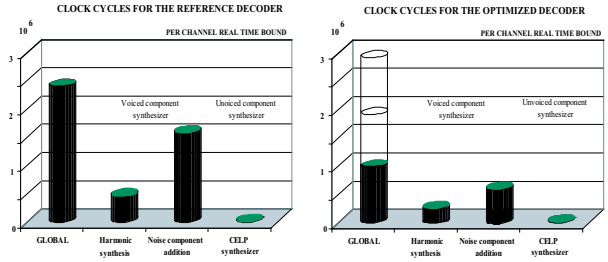

Figure 6: *Per-frame clock cycles of decoder.*

6. REFERENCES

[1] B.S. Atal M.R. Schroeder, "Code-Excited Linear Predictive (CELP)," in *Proc. IEEE ICASSP*, 1985, pp. 937–940.

[2] J.Matsumoto M.Nishiguchi, K.Iijima, "Harmonic Vector Excitation Coding at 2.0 kbps," in *IEEE Workshop on Speech Coding*, 1997, pp. 39–40.

[3] J.S.Lim D.W.Griffin, "Multiband Excitation Vocoder," in *Proc. IEEE Trans. ASSP*, 1988, pp. 1223–1235, Vol. 36.

[4] M.Nishiguchi, "MPEG-4 Speech Coding," in *The Proceedings of the AES 17^{th} International Conference*, Florence, Sept, 1999, pp. 139–146.

[5] et al N.Tanaka, "A Multi-mode Variable Rate Speech Coder for CDMA Cellular System," in *Proc. IEEE VTC* , Apr., 1996, pp. 198–202.

[6] S.Ono R.Wakatsuki M.Nishiguchi, J.Matsumoto, "Vector Quantized MBE with Simplified V/UV Division," in *Proc. ICASSP* , 1993, pp. II–151–154.

[7] J.Matsumoto M.Nishiguchi, "Harmonic and Noise Coding of LPC Residual with Classified Vector Quantization," in *Proc. ICASSP* , May, 1995, pp. I–484–487.

[8] J.Brady W.T.Adams, "Magnitude Approximations for Microprocessor Implementation," in *IEEE micro* , Oct., 1983, pp. 8,5,27–31.

[9] "www.tnt.uni-hannover.de/project/mpeg/audio/," 1998.