

Design of a variable rate algorithm for CS-ACELP coder

G. Madre, E.H. Baghious, S. Azou and G. Burel

Laboratoire d'Electronique et Systèmes de Télécommunications - UMR CNRS 6165
6, avenue Le Gorgeu - BP 809 - 29285 BREST cedex - FRANCE
e-mail : guillaume.madre@univ-brest.fr

Abstract - This paper is about the reduction of the computational complexity of the CS-ACELP codec, described in ITU recommendation G.729, and used for the transmission of voice over IP. A Voice Activity Detection module is proposed to replace the G.729 Annex B algorithm. The new procedure was developed to allow its implementation with Number Theoretic Transforms. The use of Fermat Number Transforms can reduce the cost of variable rate algorithm implementation on Digital Signal Processor (DSP).

I. INTRODUCTION

Recently, the idea of transmitting the voice on Internet, on a large scale, has appeared. Since then, Voice over Internet Protocol (*VoIP*) has been a matter of very high interest and development. But the generalization of VoIP remains currently confronted with problems of rate.

International Telecommunication Union (*ITU*) recommendation G.729 standardizes the speech coding algorithm at 8kbts/s, based on the conjugate-structure algebraic code excited linear prediction (CS-ACELP) and targeted for digital simultaneous voice and data (*DSVD*) applications [1] [2]. Its relative low complexity makes it an attractive choice for Internet telephony.

To consider the discontinuous voice activity in a conversation, a low bit rate silence compression scheme, defined in G.729 Annex B, can be employed [3]. However, to design an efficient Voice Activity Detector (*VAD*) in fixed-point arithmetic, a new procedure based on work of Aksu *et al.* [4] is introduced and implemented with Number Theoretic Transforms (*NTT*), which present the following advantages compared to Discrete Fourier Transform (*DFT*) [5] :

- They require few or no multiplications
- They suppress the use of floating point complex numbers and allow error-free computation
- All calculations are executed on a finite ring of integers, which is interesting for implementation into DSP

Hence, use of Number Theoretic Transform will reduce the delay features, by minimizing the computational complexity. The special case of Fermat Number Transforms (*FNT*), with arithmetic carried out modulo Fermat numbers, is particularly appropriate for digital computation. Its application to different functions (filter, correlation) can provide real benefits for low computational complexity.

The rest of the paper is organized as follows. In section II, we will introduce the CS-ACELP coder of ITU recommenda-

tion G.729. In a third part, we will present a variable rate algorithm. Section IV presents the concept of Number Theoretic Transform and details, more particularly, the Fermat Number Transform, which will be implemented in Voice Activity Detection procedure. In the final part, numerical results for the new algorithm are given.

II. CS-ACELP CODER G.729

The analog voice signal, sampled 8000 times per second, is taken as input signal of the coder G.729 [1]. Figure II-1 shows its principal blocks. The coder operates on frames of 10 ms. For each frame, the speech signal is analyzed to extract the parameters of the CELP model (Linear Prediction (LP) filter coefficients, adaptive and fixed codebook index), which are encoded and transmitted.

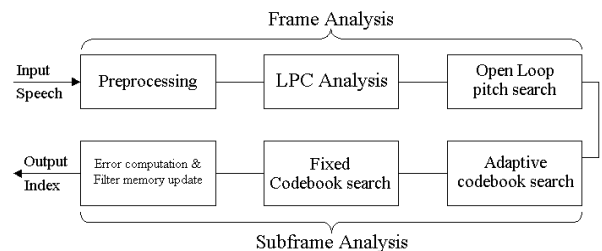


Figure II-1. Principal blocks of the coder G.729

The input signal is high-pass filtered in the preprocessing block. A 10th order linear prediction analysis yields a set of LP filter coefficients, which are converted to Line Spectrum Pairs (*LSP*) and quantized using Vector Quantization (*VQ*). The excitation signal is chosen and an open-loop pitch delay is estimated with a perceptually weighted and low-pass filtered speech signal.

The excitation parameters are determined and the gains of the adaptive and fixed codebook contributions are quantized, for subframes of 5 ms each. Finally, the filter memories are updated using the excitation signal.

III. VARIABLE RATE ALGORITHM

The problem of discriminating between speech and silence is one of the most difficult problems in speech analysis, due to large dynamic range of the speech signals and degradations of telephone lines. The standard solution to this problem is to use level tests to discriminate silence from speech and improve the decision with some measured features of the signal (energy, pitch calculation, ...).

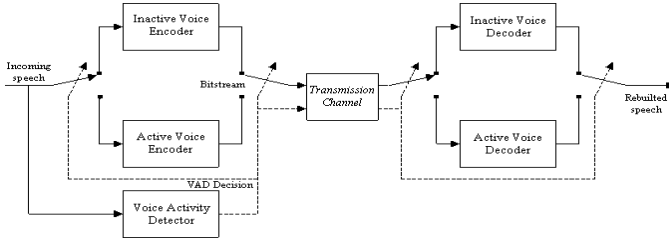


Figure III-1. Speech Communication system with VAD

Energy measurements and zero-crossings are used in VAD of G.729 annex B [3]. This algorithm performs well at high Signal-to-Noise-Ratio (SNR) levels, however its performance degrades in the presence of noise.

A. Voice Activity Detector

To replace the G.729 algorithm VAD, our study is drawn from previous results of Aksu *et al.* [4] that we adapted to CS-ACELP codec.

The VAD proposed is based on distance computations between output signals, from different filters, and references. The algorithm tracks the variation of energy levels over five sub-bands, belonging to frequency band voice [0, 4000] Hz. The frequency responses of various sub-bands [0, 500], [500, 1000], [1000, 2000], [2000, 3000] and [3000, 4000] Hz are shown in figure III-2.

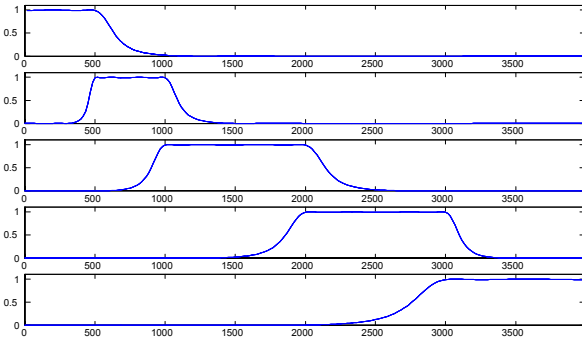


Figure III-2. Frequency Responses of the 5 sub-bands

The algorithm, detailed by the flowchart in figure III-3, requires two major steps and a smoothing decision.

Step 1 - The VAD searches two consecutive frames with gains G_{k-1} and G_k , smaller than a preset value G_a . If the condition is not met, the current frame is labelled as an ordinary CELP frame.

The gain of the k^{th} frame, of the sampled input signal s , is calculated as :

$$G_k = 10 \log_{10} \left(\sum_{i=1}^{80} s_k(i)^2 \right) \quad (1)$$

For a better comparison, G_a can be made adaptive. If the current frame is classified as a voice frame, a new reference value G_a is calculated :

$$G_a = 0.95G_a + 0.05(G_{k-1} - 5) \quad (2)$$

The initial value for G_a is equal to 40dB.

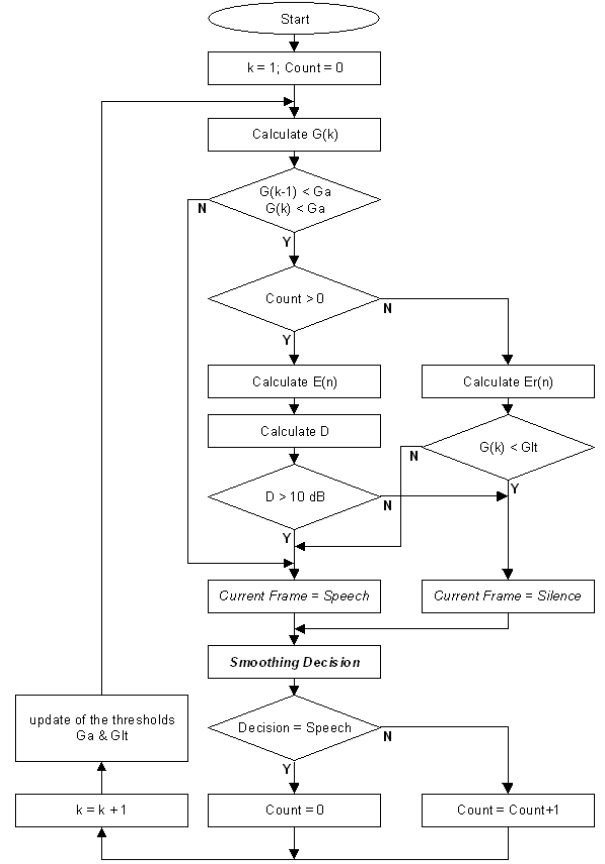


Figure III-3. Flowchart for VAD

Step 2 - If condition 1 is met, an energy parameter is calculated for the current frame as follows :

- Computation of the five outputs (denoted s_{f_n}) of sub-bands filters, whose impulse responses are noted h_n , with $n = 1, \dots, 5$. The signals s_{f_n} are obtained by a convolution $s_{f_n} = h_n \otimes s$ (\otimes denotes the convolution operator).
- Calculation of five energies :

$$E_n = \sum_{i=1}^{80} s_{f_n}(i)^2 \quad \text{with } n = 1, \dots, 5 \quad (3)$$

For each new silent frame (following a voice frame), the five energies will serve as a reference, denoted Er_n . These values are used for all the following frames of silence. In this case, the decision is made compared to a long term gain G_{lt} of the signal, with initial value equal to 20dB :

$$G_{lt} = 0.95G_{lt} + 0.05G_{k-1} \quad (4)$$

- For $count = 0$, the k^{th} frame will be labelled as "silence" if $G_k < G_{lt}$. For the following silent frame, a distance between energies is calculated as :

$$D = 10 \log_{10} \left(\sum_{n=1}^5 \frac{E_n}{Er_n} \right) - 7 \quad (5)$$

If the distance D between E_n and Er_n is smaller than 10dB, the current frame is declared as "silence".

Smoothing decision - The initial VAD decision is smoothed to avoid discontinuities in output signal. This smoothing and correction uses the value of G_{lt} too.

- In a first case, an active voice decision is extended to the current frame if the previous frame was active voice frame and the current frame energy is above G_{lt} .
- In a second case, an inactive voice decision is extended if the three previous frame are inactive voice and if the current frame energy is under G_{lt} .

B. Sub Rate Coder

If voice activity is detected, all information is transmitted (LSFs, pitch, codebook index, gains) at 8kbits/s [2]. The bit allocation of the speech coder parameters is shown in Table I.

TABLE I
BIT ALLOCATION OF THE CODER G.729

Parameters	Subframe		Total per frame
	n°1	n°2	
LSFs	1+7+5+5		18
Adaptive Codebook Delay	8	5	13
Pitch Delay Parity	1	-	1
Fixed Codebook Index	13	13	26
Fixed Codebook Sign	4	4	8
Codebook Gains	3+4	3+4	14
Total			80

For silent frame, LSF and gain information is transmitted only for the first frame. The LSF index, obtained by two-stage vector quantization, is represented with 13 bits. The gain is scalar quantized using 7 bits nonuniform quantizer.

These parameters, in the decoder, are repeated for the rest of the region labeled as "silence" unless the variation of energy, between two consecutive silent frames, is significant. In this case, the gain is transmitted again for a better characterization of the "comfort noise".

IV. NUMBER THEORETIC TRANSFORM

To develop the previous VAD and CS-ACELP codec in fixed point arithmetic with low computational complexity, we propose to use Number Theoretic Transforms (NTT).

An NTT [5] presents the same form as the DFT but is defined over finite rings. All arithmetic must be carried out modulo M . The modulo M may be equal to a prime number or to a multiple of primes, since NTTs are defined over Galois Field ($GF(M)$). The N^{th} root of the unit in \mathbb{C} , $e^{j\frac{2\pi}{N}}$, is replaced by the N^{th} root of the unit in $GF(M)$ represented by the term $\langle \alpha \rangle_M$, where $\langle \cdot \rangle_M$ denotes the modulo M operation.

An NTT of a discrete time signal x and its inverse are given respectively by :

$$X(k) = \left\langle \sum_{n=0}^{N-1} x(n)\alpha^{nk} \right\rangle_M \quad (6)$$

$$x(n) = \left\langle N^{-1} \sum_{k=0}^{N-1} X(k)\alpha^{-nk} \right\rangle_M \quad (7)$$

with $n, k = 0, 1, \dots, N - 1$. α represents the generating term and N the length of the transform. N being a prime number, there exists an integer N^{-1} such $\langle N.N^{-1} \rangle_M = 1$.

Different conditions must be satisfied for an NTT to exist over a Galois Field $GF(M)$ [6] [7] :

- $\gcd(\alpha, M) = \gcd(N, M) = 1$ i.e. $\langle \alpha^N = 1 \rangle_M$
- $N | \gcd(p_i - 1, p_j - 1), \forall (i, j \neq i) \in [1, k]^2$ for $M = \prod_{i=1}^k p_i$
- $\gcd((\alpha^i - 1), M) = 1, \forall i \in [1, N - 1]$

(gcd is the greatest common divisor and $a|b$ means that the remainder of $\frac{a}{b}$ is equal to zero)

For an efficient implementation of Number Theoretic Transform on processor, the choice of parameters is important. If possible, the values N and α are chosen as a power of 2 to allow replacement of multiplications by bit shifts.

A. Cyclic Convolution Property

The Number Theoretic Transforms, originally developed for rapid computation of convolution, have all the Cyclic Convolution Property (CCP) [6] :

$$U \otimes V = T_N^{-1} \{T_N(U) \bullet T_N(V)\} \quad (8)$$

where U and V represent the two sequences whose convolution is desired, T_N an NTT of length N and T_N^{-1} its inverse. The two operators \otimes, \bullet denote the convolution and the term by term multiplication, respectively.

B. Fermat Number Transform

The choice of a modulo equal to a Fermat number [7], $F_t = 2^{2^t} + 1$ with $t \in \mathbb{N}$, offers numerous possibilities for length N of the transform (see Table II). The values of N and α associated to a Number Theoretic Transform, defined for a modulo F_t , are given by $N = 2^{t+1-i}$ and $\alpha = 2^{2^i}$ with $i < t$. The NTT defined over the Galois Field $GF(F_t)$, is called Fermat Number Transform (FNT) :

$$X(k) = \left\langle \sum_{n=0}^{2^{t+1-i}-1} x(n)2^{2^i nk} \right\rangle_{F_t} \quad (9)$$

$$x(n) = \left\langle -2^{2^{t-i}-1-(t-i)} \sum_{k=0}^{2^{t+1-i}-1} X(k)2^{-2^i nk} \right\rangle_{F_t} \quad (10)$$

with $k, n = 0, 1, \dots, 2^{t+1-i} - 1$.

TABLE II
POSSIBLE COMBINATIONS OF PARAMETERS OF AN FNT

t	modulo	N	N
	F_t	for $\alpha = 2$	for $\alpha = \sqrt{2}$
1	$2^2 + 1 = 5$	4	-
2	$2^4 + 1 = 17$	8	16
3	$2^8 + 1 = 257$	16	32
4	$2^{16} + 1 = 65537$	32	64
5	$2^{32} + 1$	64	128
6	$2^{64} + 1$	128	256

A Fermat Number Transform satisfies the Cyclic Convolution Property and requires about $N \log_2 N$ simple operations (bit shifts, additions) but no multiplication, while a DFT requires a number of multiplications of about $N \log_2 N$. Moreover, this transform admits a fast NTT-type computational structure (butterfly implementation).

C. Convolution analysis

The convolution computation of two sequences of N samples, using the Cyclic Convolution Property, requires the use of three NTTs of length $2N$ (the sequences are extended by zeros). To avoid zeros addition, an algorithm is summarized below [8].

To calculate the correlation of x and h , two sequences of length N , the following procedure is applied :

- First, new sequences are formulated from x :

$$\begin{cases} x_1(k) = \begin{cases} x(k) & 0 \leq k < \frac{N}{2} \\ 0 & \frac{N}{2} \leq k < N \end{cases} \\ x_2(k) = \begin{cases} 0 & 0 \leq k < \frac{N}{2} \\ x(k) & \frac{N}{2} \leq k < N \end{cases} \end{cases} \quad (11)$$

- Two other sequences, denoted h_1 and h_2 , are in the same way defined from h , with $k = 0, \dots, N - 1$.
- Let X_1, X_2, H_1 and H_2 , be the respective NTTs of x_1, x_2, h_1 and h_2 . Then, X, H are determined as :

$$\begin{cases} X(k) = X_1(k) + X_2(k) \\ H(k) = H_1(k) + H_2(k) \end{cases} \quad (12)$$

- Now, define three new sequences :

$$\begin{cases} U(k) = H(k)X(k) \\ V(k) = H_1(k)X_1(k) \\ W(k) = H_2(k)X_2(k) \end{cases} \quad (13)$$

- Finally the convolution $y = x \otimes h$ is obtained from the inverse NTTs of U, V and W :

$$\begin{cases} y(n) = v(n) & 0 \leq n < \frac{N}{2} \\ y(n) = u(n) - w(n) & \frac{N}{2} \leq n < N \\ y(n) = u(n - N) - w(n - N) & N \leq n < \frac{3N}{2} \\ y(n) = w(n) & \frac{3N}{2} \leq n < 2N \end{cases} \quad (14)$$

V. NUMERICAL RESULTS

Numerical simulations have been conducted to evaluate the performances of the presented VAD. A comparison with the standard VAD, described in G.729 annex B, has also been realized under different environmental conditions.

In these tests, we compare the decisions to a database which was hand-labeled as voice or silence.

Averaged results are computed as a function of SNR, across a Monte Carlo simulation consisting of 3000 runs, for a same sentence pronounced by a female and a male talker. The noise is additive, white and gaussian.

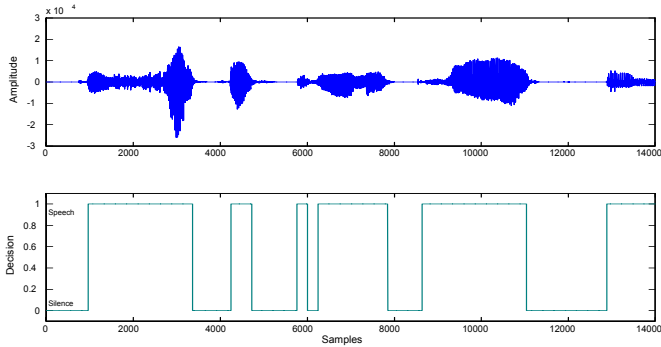


Figure V-1. Female Talker (upper plot)
Silence/Voice Frames (lower plot)

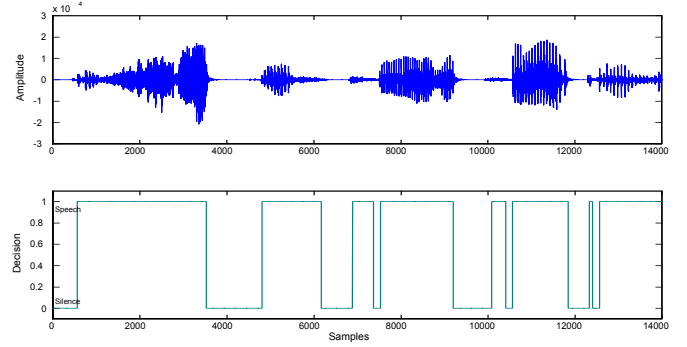


Fig. V-2. Male Talker (upper plot)
Silence/Voice Frames (lower plot)

As shown by Figures V-1 and V-2, the right decisions the VADs are expected to do, for recording extracts without noise, are known in advance.

A. Performances Comparisons

Both VAD algorithms give similar performances for clean speech or high SNR. On the other hand, as the SNR decreases, the difference between the VADs becomes more pronounced. In Figures V-3 and V-4, the dotted curves represent the standard deviation of the results around the average values shown by the solid curves. The dashed lines correspond to the misclassification percentage observed and to the theoretical rate for nondisturbed signals.

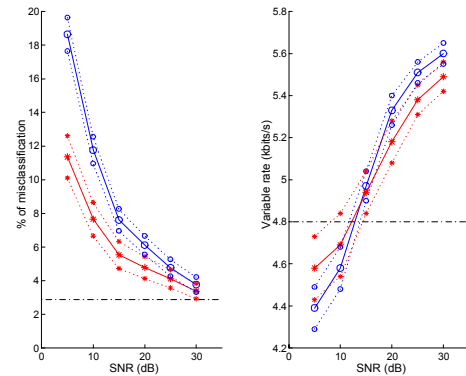


Figure V-3. Comparisons of VAD for the female voice
○ G.729 B VAD * Proposed VAD

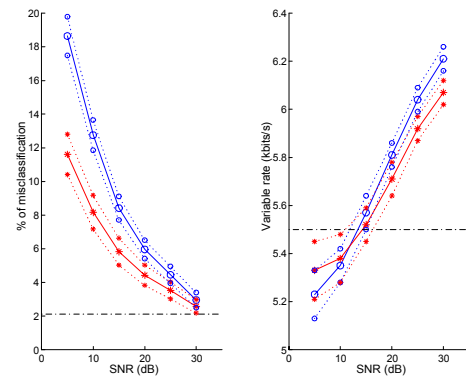


Figure V-4. Comparisons of VAD for the male voice
○ G.729 B VAD * Proposed VAD

The rate evolutions can be explained by the misclassification increases of both VADs. The less the number of voice frames are detected, the more the rate decreases. For SNRs lower than 10 dB, the rates are nonsignificant of VAD performances.

B. Implementation of FNT-based VAD

Various filtering involved in the described VAD are easily implemented using FNTs. The length of the five sub-band filters impulse responses, illustrated by Figure V-5, has been chosen equal to 128, with samples 16 bits quantized.

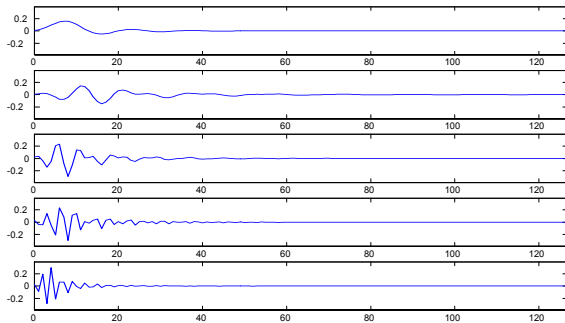


Figure V-5. Impulse Responses of the five sub-bands filters

Comparing with a DFT implementation, the proposed FNT solution, using the convolution computation presented previously, requires a reduced number of multiplications to obtain the five filter outputs. For each filter processing, a frame admitting gains $\{G_{k-1}, G_k\}$ smaller than threshold G_a , the FNT algorithm requires about 400 multiplications against 3100 for a DFT implementation.

In the other functions of VAD, the computational complexity of floating point code is slightly smaller, but comparable, than 32 bits fixed-point arithmetic algorithm (a table for logarithm operations, already existing in the coder G.729, is used) [9].

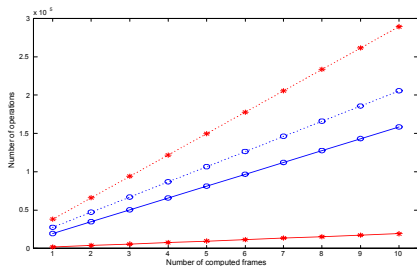


Figure V-6. Operations required for the 5 filters
 o Algorithm with DFT * with FNT

The computation gain is more significant for the filtering process. In Figure V-6, the operations number, required for the five filters, is represented in dotted lines for simple operations (additions, bit shifts) and in solid lines for multiplications.

VI. CONCLUSION

In this paper, an efficient implementation of a variable rate algorithm for CS-ACELP codec is proposed. The input speech frames are classified into voice and silence parts by a novel VAD algorithm, which replaces the G.729 Annex B algorithm. This robust VAD tracks the variation of energy levels of five spectral

sub-bands and makes a decision with moderate complexity. Following stages would be to make similar tests for different types of noise (nonstationary noise) and to evaluate the performances of both VADs comparing the subjective speech quality, in the form of Mean Opinion Score (MOS).

Although the selected application is the CS-ACELP coder G.729, the proposed silence detection procedure can be adapted to other types of coder using a VAD.

To limit the cost implementation of the variable rate procedure, Number Theoretic Transforms have been used. In particular, it is shown that Fermat Number Transform reduces the computational complexity of different algorithms involved in the VAD procedure.

Moreover, the FNT implementation could benefit to other functions of associated speech coder (autocorrelation [10], LP coefficients computation [9], filters, ...).

REFERENCES

- [1] ITU-T Recommendation. G.729, "Coding of Speech at 8 kbits/s using Conjugate Algebraic Code-Excited Linear Prediction (CS-ACELP)", June 1995.
- [2] R. Salami, C. Laflamme, B. Bessette, J.P. Adoul, "ITU-T G.729 Annex A : Reduced complexity 8kbits/s CS-ACELP codec for digital simultaneous voice and data", IEEE Communications Magazine, vol. 35, pp. 56-63, September 1997.
- [3] A. Benyassine, E. Shlomot, H.-Y. Su, "ITU-T recommendation G.729 Annex B : A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications", IEEE Communications Mag., vol. 35, pp. 64-73, September 1997.
- [4] E.B. Aksu, A.E. Ertan, H.G. Ilk, H. Karci, O. Karpat, T. Kolkak, L. Sendur, M. Demirekler, E. Cetin, "Implementation of a Variable-Bit Rate MELP Vocoder on TMS320C548", 2nd European DSP Education and Research Conference, Noisy-le-Grand, September 1998.
- [5] G.A. Julien, "Number Theoretic Techniques in Digital Signal Processing", Book Chapter Advances in Electronics and Electron Physics, Academic Press Inc., vol. 80, Chapter 2, pp. 69-163, 1991.
- [6] R. Blahut, "Fast algorithms for digital signal processing", Addison-Wesley Publishing Company, 1985.
- [7] R.C. Agarwal, C.S. Burrus, "Fast convolution using Fermat number transform with application to digital filtering", IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-22, N^o2, pp. 87-97, 1974.
- [8] W. Shu, Y. Tianren, "Algorithm for linear Convolution using Number Theoretic Transforms", Electronics Letters, vol. 24, N^o5, March 1988.
- [9] A.V. Aho, J.E. Hopcroft, J.D. Ullman, "The design and analysis of computer algorithms", Addison-Wesley Publishing Company, 1974.
- [10] S. Xu, L. Dai, S.C. Lee, "Autocorrelation Analysis of Speech Signals Using Fermat Number Transform", IEEE Trans. on Signal Proc., vol. 40, N^o8, August 1992.