FACTA UNIVERSITATIS (NIŠ)
SER.: ELEC. ENERG. vol. 17, December 2004, 389-402

A New Speech Compression Method

Miranda Nafornita, Alexandru Isar, and Dorina Isar

Abstract: In this paper a new speech compression method is presented. The traditional speech compression method is based on linear prediction. The compression method, proposed in this paper, is based on the use of an orthogonal transform, the discrete cosine packets transform. This method is well suited for the speech processing, taking into account the sine model of this kind of signals and because this transform converges asymptotically to the Karhunen-Loève transform. After the computation of the discrete cosine packets transform, the coefficients obtained are processed with a threshold detector, who keeps only the coefficients superior to a given threshold. This way the number of non zero coefficients is reduced doing the compression. The next block of the compression system is the quantization system. This is build following the speech psycho-acoustic model. The proposed compression method is transparent, the compression rate obtained is important and the operations number and the memory volume used are not very high.

Keywords: Speech compression, cosine packets, adaptivequantization, perceptual thresholding.

1 Introduction

On the basis of the classical papers written by Shannon, [1] and Kolmogorov, [2], recently was highlighted a strong connection between the systems proposed in many lossy compression standards and the harmonic analysis, [3]. All these systems uses orthogonal transforms and quantizers. The architecture of a data compression system based on the use of an orthogonal transform is presented in Fig. 1. This system will be analyzed in the following for the case of the Discrete Cosine Packets, DCPT, orthogonal transform. It will be proved that this transform is well selected for the speech compression. The most important signals in Fig. 1 are:

Manuscript received June 22, 2004.

The authors is are with Electronics and Telecommunications Faculty, "Politehnica" University, 2 Bd. V. Parvan, 1900, Timisoara, Romania (email: isar@etc.utt.ro).

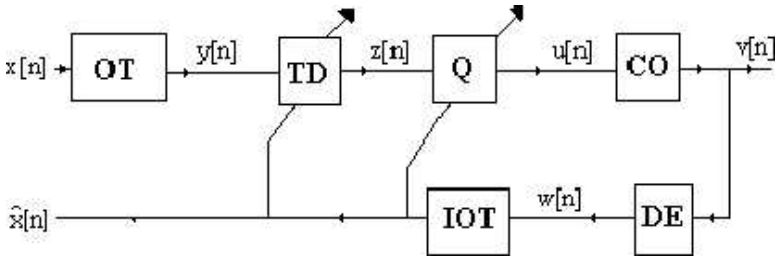

Fig. 1. A data compression architecture. OT is Orthogonal transform computation block, TD is Threshold Detector (a system that rejects all the small samples inferior to a given threshold), Q is Quantizer, CO is Loss less Compression System, DE is Decoder, realizes the inverse operation of CO and IOT is Inverse Orthogonal Transform computation block, realizes the inverse operation of OT.

the signal to be processed, $x[n]$, its compressed version $v[n]$, and the reconstructed signal, $\widehat{x}[n]$. The samples of $x[n]$ are correlated. This means that a fraction of the information carried by each sample is contained in its neighbors. This is the reason why rejecting a sample, the information contained in its neighbors is also affected. The role of the orthogonal transform is to de-correlate the signal to be compressed. After its computation, a new signal is obtained. The dependency of the informational content of each sample of this new signal, $y[n]$, on the informational content of its neighbors is wicker. Hence rejecting a sample, the informational content of its neighbors is less affected after the computation of the orthogonal transform. The rejection of a sample produces a smaller loss of information if its magnitude is smaller. This is the reason why the threshold detector is present in Fig. 1. The structure of this paper is the following. In section 2 are presented three discrete wavelet transforms. A statistical analysis of each of them is presented. All these transforms asymptotically converge to the Karhunen-Loève transform. At the end of this section is explained why the DCPT is a good candidate for the speech compression. In section 3, an adaptive threshold detector is described. In section 4 is described a quantizer. The other blocks of the system in Fig. 1 are also described. The aim of section 5 is to present some simulation results. In section 6 are presented some conclusions.

2 Discrete Wavelet Transforms

In the following a statistical analysis of three discrete wavelet transforms: the discrete wavelet transform, DWT, [4], the discrete wavelet packets transform DWPT and the DCPT, [5], is presented. This analysis is based on simulations. The correspondent transform of a realization of a colored noise is computed. The result is practically a white noise. For the first two transforms, the mathematical proofs can

be found in [6]. For the DWT a similar proof can be found in [7]. For the third transform, a statistical analysis is reported in [8].

2.1 The DWT case

The result is presented in Fig. 2. In the top of this figure is presented the power spectral density of the input noise. This a colored noise, generated by filtering a white noise. The filter used was a running averager with 20 tapes. In the bottom of Fig. 2 is presented the power spectral density of the result obtained, computing the DWT of the realization with the power spectral density represented in top. Because the envelope of the power spectral density represented in bottom is practically constant we can assert that the output signal is a white noise. So, the DWT asymptotically converges to the Karhunen-Loève transform.

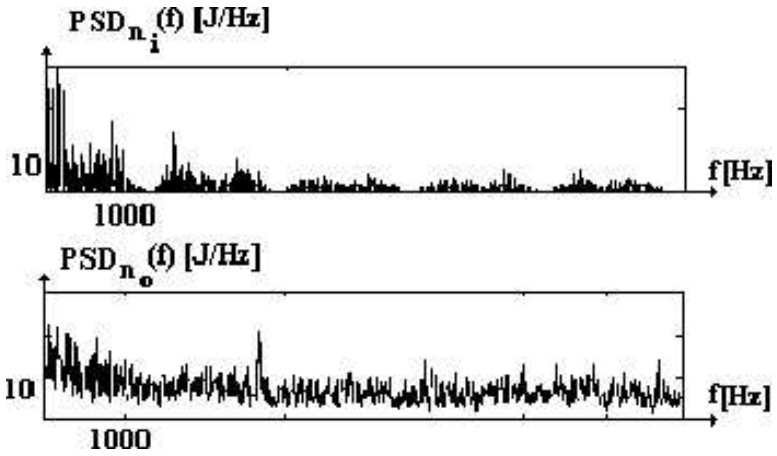

Fig. 2. The whitening effect of the DWT.

2.2 The DWPT case

In Fig. 3 is illustrated the DWPT whitening effect. The experiment already described is repeated for the case of the DWPT. In top of Fig. 3 is represented the power spectral density of the input signal. This is a realization of a colored noise obtained by band-pass filtering a white noise. In bottom is represented the power spectral density of the signal obtained after the computation of the DWPT of the signal with the power spectral density represented in top.

Because the envelope of the power spectral density is constant almost everywhere, like the power spectral density of a white noise, we can assert that the DWPT also converges to the Karhunen-Loève transform.

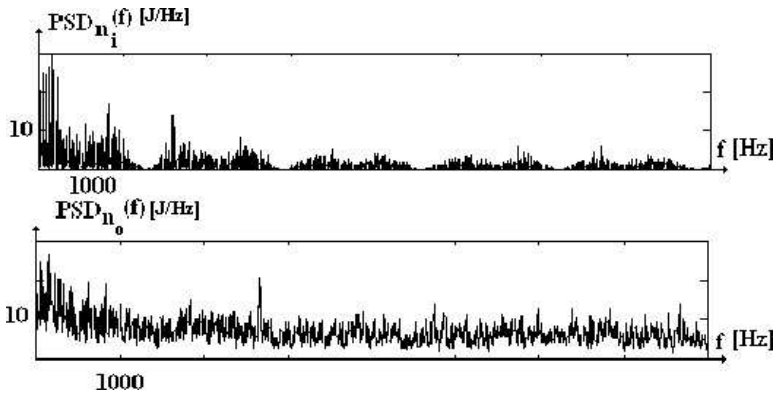

Fig. 3. The whitening effect of the DWPT.

2.3 The DCPT case

The DCPT is a combination of the discrete cosine transform, DCT, with the theory of wavelet packets. In the case of this transform the segmentation of the support of the signal to be analyzed (considered of length N) is realized. The lengths of the segments obtained depend on the number of iterations of the transform performed. For the iteration index m these blocks have a length of $2^{-m}N$. The signal contained in each such segment is transformed using the DCT. The asymptotic analysis of DCPT is based on the observation that for $N \to \infty$, the DCT asymptotically converges to the Karhunen-Loève transform, [8]. If $N \to \infty$, the number of samples of each DCT coefficients sequence, corresponding to a given segment, at a given iteration, tends to infinity and the coefficients of the corresponding DCT converge to the Karhunen-Loève transform. Hence, the DCPT asymptotically converges to the Karhunen-Loève transform. The DCPT whitening effect is presented in Fig. 4. In top is represented the power spectral density of a colored noise, obtained filtering white noise with the aid of a running averager and in bottom is represented the power spectral density of the signal obtained after the computation of the DPCT of the signal with the power spectral density represented in top. The envelope of the power spectral density of the signal represented in bottom is a good approximation of a constant. Hence the signal with the power spectral density represented in bottom is practically a white noise.

Hence the whitening effect of the DCPT was highlighted. Comparing the figures 2, 3 and 4, it can be observed that the DCPT has the higher convergence rate to a white noise. It is followed by the DWT. The DWPT has the slower convergence. Each transform can be used in a compression system because all of them converge to the Karhunen-Loève transform. Taking into account the convergence speed the best is the DCPT. This transform has also another advantage, the temporal localiza-

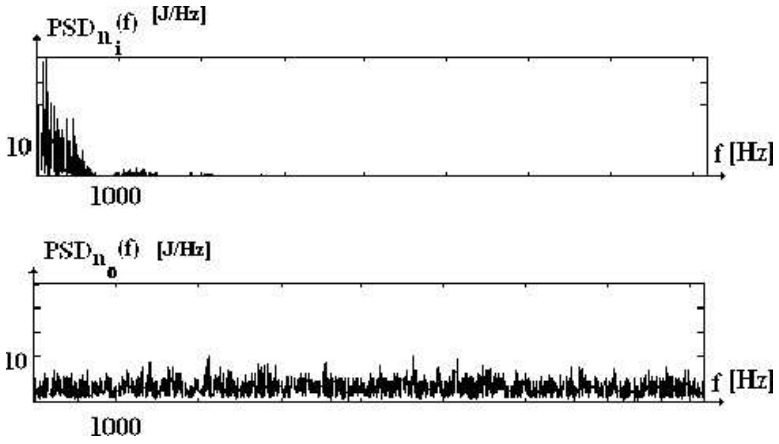

Fig. 4. The whitening effect of DCPT.

tion of the analysis filters. Also it is recommended for the compression of speech taking into account the sinusoidal model of the speech. This proposition is proved in the following.

2.4 Selecting the best mother wavelets

The quality of a compression system can be appreciated with the aid of his rate-distortion function. A compression system is better than another if, at equal distortions, it realizes a higher compression rate. The maximization of the compression rate can be done, if a good selection of the parameters of each block represented in Fig. 1 is performed. In the following is analyzed the first block. One of the three orthogonal transforms, already mentioned, and its parameters must be selected, taking into account the fact that the signal to be processed is a speech signal. Each of the three orthogonal transforms already mentioned has some parameters. A common parameter for all the three transforms is the mother wavelets. This parameter must be selected in accordance with the type of the signal to be compressed. Each spoken word is a sequence of tons having different intensities, frequencies and durations. Every ton is a sinusoidal signal with specific amplitude, frequency and duration. This is the speech sinusoidal model. Its mathematical description is, [9]

$$x(t) = \sum_{q=1}^{Q(t)} A_q(t) \cos \theta_q(t) \tag{1}$$

where the components are named partials. Each term of this sum is a double modulated signal. So, the speech signal is not stationary. But, usually the speech signal is regarded like a sequence of stationary signals. Segmenting the speech signal in

blocks with duration inferior to 25 ms, a sequence of stationary signals is obtained. On every segment the speech model is of the form

$$x_s(t) = \sum_{q=1}^{Q} A_q \cos \omega_q t. \tag{2}$$

This decomposition is very similar to a decomposition of the signal $x_s(t)$ into a cosine packet basis.

The decomposition of the same signal into an orthogonal wavelet basis is, [10]

$$x_s(t) = \sum_{k=1}^{K} \sum_{l=1}^{L} \left\langle x_s(t), \psi_{k,l}(t) \right\rangle \psi_{k,l}(t) \tag{3}$$

where $\psi_{k,l}(t)$ are the wavelets generated by the mother wavelets $\psi(t)$. The compression rate obtained using a specified mother wavelets is higher if the number of not null coefficients,

$$d_{k,l} = \left\langle x_s(t), \psi_{k,l}(t) \right\rangle \tag{4}$$

N_ψ is smaller. But

$$d_{k,l} = \int_{-\infty}^{\infty} x_s(t) \cdot \psi_{k,l}^*(t)\, dt = r_{x_s,\psi_{k,l}}(0) \tag{5}$$

where the right hand side member represents the value of the correlation of $x_s(t)$ and $\psi_{k,l}(t)$, computed in zero. This correlation measures the similarity of the two signals. So, the value of the coefficient $d_{k,l}$ is higher if the signals $x_s(t)$ and $\psi_{k,l}(t)$ are more similar.

But the energy of the signal $x_s(t)$ can be computed using the relation

$$E_x = \sum_{k=1}^{K} \sum_{l=1}^{L} \left| d_{k,l} \right|^2. \tag{6}$$

Because the signal energy is a constant, not dependent on the mother wavelets used, it can be said that the number N_ψ is smaller if the magnitude of not null coefficients is higher. But this magnitude is higher, (see (5)), if the signals $x_s(t)$ and $\psi_{k,l}(t)$ are more similar. Following (2) the higher similarity is obtained when the functions $\psi_{k,l}(t)$ are elements of a cosine packet. This is the reason why, for the compression of speech signals, the best transform is the DCPT. Another parameter of the DCPT is the cost functional used for the selection of the best packet. This transform is an adaptive one. The result of its utilization in a given application can

be optimized using the best packet selection procedure. This is a very efficient procedure able to enhance very much the quality of a given signal processing method. There are some cost functions that can be minimized for the selection of the best cosine packet. The most used is the entropy but its utilization do not realizes the maximization of the compression rate. The optimal cost functional for compression is that realizing the minimization of the number of coefficients superior to a given threshold, t, (who determines the distortion), N_s. Using this cost functional, N_s coefficients superior to the threshold t are obtained. So, at the output of the TD, from Fig. 1, N_s not nulls coefficients are obtained. This is a minimal number because it was obtained using the appropriate cost functional for the selection of the best packet. This is the reason why this cost functional realizes the maximization of the compression rate. Increasing the threshold value t, the number N_s decreases and the compression rate increases. Unfortunately the reconstruction distortion increases when t increases. This is the reason why the threshold's value t must be controlled to assure the compression transparency. Hence, the block TD must be an adaptive one. Another parameter of the DCPT who must be considered for the optimization of the compression is its number of iterations.

3 The Threshold Detector

One of the most important block for the system in Fig. 1 is the threshold detector. It rejects all the coefficients of the orthogonal transform inferior to a given threshold. This is the mechanism doing the compression. It is an adaptive system. Its input-output relation is

$$z[n] = \begin{cases} y[n], & |y[n]| > t \\ 0, & |y[n]| \leqslant t. \end{cases} \tag{7}$$

Analyzing the system in Fig. 1 it can be observed that the distortion due to the compression has the mean squared value

$$D = E\left\{ \left(x[n] - \widehat{x}[n] \right)^2 \right\}.$$

Because the DCPT and its inverse the IDCPT are orthogonal transforms, the last relation becomes

$$D = E\left\{ \left(y[n] - u[n] \right)^2 \right\}. \tag{8}$$

Two blocks of the system from Fig. 1 are responsible for this error, the threshold detector and the quantizer. The threshold value, t and the architecture of the quantization system must be selected to satisfy the following condition

$$D < \alpha \cdot E_x, \qquad \alpha < 1 \tag{9}$$

where E_x represents the energy of the input signal $x[n]$.

4 The Quantizer

Another very important block for a compression system is the quantizer. For the compression of speech the quantizer structure must take into account the particularities of this signal. These particularities are highlighted by the psycho-acoustic model of the speech. A non-uniform quantization, based on the psycho-acoustic model, [11], can be used. The most important disadvantage of such a method is the high computational volume, the computation of the masking threshold demanding a large amount of operations. In the following a very simple solution of non-uniform quantization is proposed. The sequence obtained at the output of the TD in Fig. 1, represents the instantaneous spectrum of the current segment. This is the reason why a perceptual quantization, based on the psycho-acoustic model, can be performed. The support of the signal obtained at the output of TD is divided into 32 intervals of equal length, containing the signals $z_k[n]$. These intervals correspond to the critical bands of the psycho-acoustic model. For each band, the block TD eliminated all the spectral components with the magnitude inferior to a given threshold, t. This value is an approximation of the masking threshold. The uniform quantization of each block is performed. To do this quantization, the maximal values of the signals $z[n]$ and $z_k[n]$, z_M and z_{kM} are detected. For each band a given number of bits is allocated. This allocation procedure is based on the values z_{kM} . For every value z_M ,6 bits are allocated, (2^6 quantization levels). For each value z_{kM}, a number of

$$\gamma_k = [[\frac{z_{kM}}{z_M} \cdot 2^6]] \tag{10}$$

quantization levels is allocated, where $[[\cdot]]$ represent the entire part. The quantization of the corresponding interval is realized using the transformation

$$u_k[n] = [[\frac{z_k[n]}{z_{kM}} \cdot \gamma_k]]. \tag{11}$$

So, a level normalization, in each block, is performed. The corresponding denormalization will be performed by the block DE. The good performances of this quantization procedure are based on the de-correlation property of the DCPT. Due to this property a lot of the values z_{kM} are small. The corresponding values γ_k are zeroes. This is the reason why the total number of bits, allocated to the samples of the signal $u[n]$, is very small versus the number of bits of the signal $x[n]$.

5 Computing the Reconstruction Distortion

In the following a superior bound of the reconstruction distortion is computed. This distortion is given in the relation (8). It has two components, produced by the

block TD and by the quantization block. A superior bound of the reconstruction distortion, obtained after the adaptive compression based on the DCPT is $N \cdot t^2$ where N represents the processing signal samples number and t is the value of the threshold.

Proof.

The mean square approximation error of the signal $y[n]$ with the signal $u[n]$ is

$$\varepsilon = E\left\{ (y[n] - u[n])^2 \right\} = \sum_{k=1}^{K} y^2 \left[n_k\right] + \varepsilon_1 \tag{12}$$

where n_k represent the position of samples with the magnitude inferior to t, in the sequence $u[n]$. Let K be the number of those samples. Let $o[n]$ be the signal obtained after the ordering of the samples of the signal $y[n]$ following the order of their magnitudes. The mean squared error becomes

$$\varepsilon = \sum_{k=1}^{K} o^2[k] + \varepsilon_1 \leq K \cdot t^2 + \varepsilon_1 \tag{13}$$

where the first term represents the error introduced by the TD. The mean squared error produced by the quantizer is

$$\varepsilon_1 = \sum_{n=1}^{N} (z[n] - u[n])^2 \qquad \text{for} \qquad z[n] \neq 0. \tag{14}$$

After the segmentation of the support of the signal $z[n]$ in 32 bands it is obtained

$$\varepsilon_1 = \sum_{k=1}^{32} \sum_{p_k=1}^{N_k} \left(z_k\left[p_k\right] - u_k\left[p_k\right]\right)^2$$

where p_k is the index for the samples corresponding to the k'th band. It is supposed that there are N_k such samples. Let q_k be the quantization step for the k'th band. Because the difference $z_k\left[p_k\right] - u_k\left[p_k\right]$ is inferior to the value q_k, it can be written

$$\varepsilon_1 \leq \sum_{k=1}^{32} N_k \cdot q_k^2.$$

Hence

$$\varepsilon \leq \sum_{k=1}^{32} N_k \cdot q_k^2 + K \cdot t^2.$$

If all the quantization steps are inferior to the value of threshold, t, then the last relation can be written in the following form

$$\varepsilon \leq \sum_{l=K+1}^{N} t^2 + K \cdot t^2 = (N - K) \cdot t^2 + K \cdot t^2 = N \cdot t^2. \tag{15}$$

The proposition was proved.

So, to keep the distortion D under the value αE_x, (E_x represents the energy of the signal $x[n]$) it is sufficient to select the value of the threshold

$$t = \sqrt{\frac{\alpha E_x}{N}}. \tag{16}$$

The constant α can be written using the signal to noise ratio of the signals $u[n]$ and $\widehat{x}[n]$ (who have the same energy), rsb_0. It can be written

$$rsb_0 = 10 \cdot \log_{10} \frac{E_x}{D} \geqslant -10 \cdot \log_{10} \alpha. \tag{17}$$

So, an inferior bound of the signal to noise ratio, depending on α, was established

$$\beta = -10 \cdot \log_{10} \alpha.$$

Taking the sign equal in the relation (17), an inferior bound for the constant α can be obtained

$$\alpha_m = 10^{-\frac{\beta}{10}}$$

Using this value and the relation (16) an inferior bound of the threshold t can be obtained

$$t_m = \sqrt{10^{-\frac{\beta}{10}} \cdot \frac{E_x}{N}}. \tag{18}$$

Hence, selecting for the threshold a value t superior to t_m, a value of the output signal to noise ratio superior to β is obtained. Unfortunately the exact value of the output signal to noise ratio, rsb_0 can not be known in advance. This is the reason why an adaptive algorithm for the threshold selection is recommended. This algorithm can use the value t_m for initialization. The threshold value is increased starting with this value. At each iteration the output signal to noise ratio, rsb_0, is computed. If this value is superior to β, the threshold value increasing process is continued. The algorithm stops when for the first time the value rsb_0 becomes inferior to β.

6 Other Blocks of the Compression System

The use of the coder CO, from Fig. 1, realizes an increasing of the compression rate without affecting the reconstruction distortion level because this block produces a loss less compression. The construction of this block uses one of the classical compression algorithms like for example the Huffman algorithm or the arithmetic coding.

The signal $v[n]$ represents the result of the compression procedure. The other two blocks of the system in Fig. 1 are used in the reconstruction phase. The block DE realizes the decoding of the signal $v[n]$. At its output the signals $u_k[n]$ and the sequence z_{kM}, $k = \overline{1,32}$, are obtained. Using the relation (10), the values γ_k are computed. After that, the denormalization

$$w_k[n] = \frac{u_k[n]}{\gamma_k} \cdot z_{kM} \tag{19}$$

is performed. Concatenating the signals $w_k[n]$,the signal $w[n]$ is obtained. The last block in Fig. 1 computes the IDCPT. The result is the signal $\widehat{x}[n]$. This signal represents the result of the reconstruction procedure. Using this signal, the distortion D can be computed. All the operations already described are repeated, for different threshold values, to maximize the global compression rate, under the constraint that rsb_0 to be superior to β.

7 Simulation Results

In the following a simulation example is presented. The proposition to be compressed is: "Huston we have a problem". The source is in wave format. The sampling frequency used to obtain this file is of approximative 44 KHz, each sample being coded on 16 bits. The waveform of this signal is presented in the following figure.

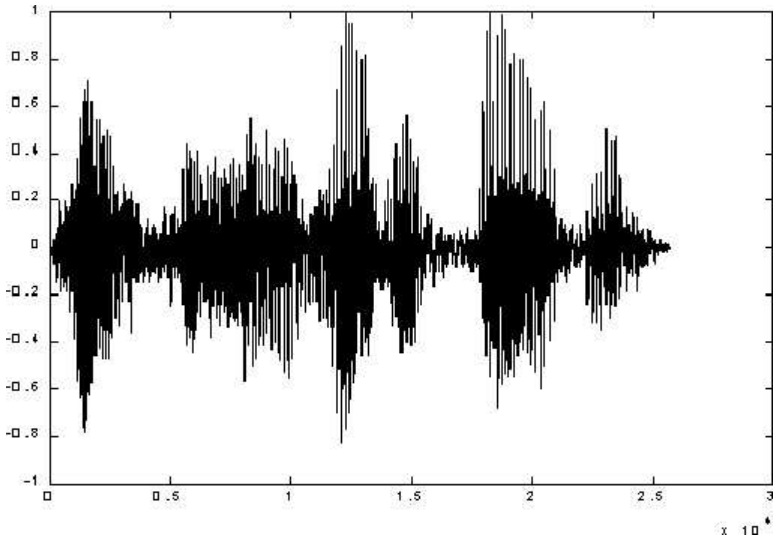

Fig. 5. The waveform of the signal to be compressed.

This signal was segmented in blocks containing 1024 samples (the duration of each block being inferior to 25 ms).

To limit the bits number required for the best packet tree coding, the maximal number of the DCPT iterations was fixed to three. In the following table the results obtained are presented. The 25 segments obtained after the reconstruction

Table 1. Some simulation results.

No.	Compression rate	rsb_0 $[dB]$	Iterations number
1	7.40	19.41	3
2	15.99	18.98	0
3	12.66	18.99	3
4	4.43	19.11	2
5	4.11	18.83	0
6	7.87	19.53	2
7	13.39	19.43	0
8	12.79	19.33	0
9	16.43	19.65	2
10	11.34	19.35	2
11	8.12	19.50	2
12	6.18	19.28	3
13	7.82	19.59	2
14	12.68	19.76	0
15	18.31	19.79	0
16	7.48	18.53	1
17	4.05	16.34	0
18	18.04	19.37	3
19	12.11	19.83	0
20	8.35	19.11	0
21	9.33	19.77	3
22	14.19	19.09	1
23	14.21	18.20	1
24	14.83	19.28	1
25	8.53	17.45	1

(realized on each segment) were concatenated obtaining the reconstructed signal after the compression. Analyzing the table, the good performances of the proposed compression method (compression rate and output signal to noise ratio) can be observed. The smallest compression rate, 4.05, was obtained on the 17'th segment and the better compression rate, 18.3, was obtained on the 15'th segment. The output signal to noise ratios distribution is uniform. The smallest value, 16.33 dB, was obtained for the 17'th segment and the best value, 19.83, on the 19'th segment. All these values are high enough to certify a good quality reconstruction. The mean

compression rate is of 10.82. This is a sufficiently high value, taking into account the fact that any loss less compression method was not used (the blocks CO and DE were not simulated).

8 Conclusion

A new lossy compression system, with the control of the losses, is proposed. Its construction is based on the rate-distortion function concept. Taking into account the high redundancy of the speech signal, such a method is very appropriate. Using this method, a mean compression rate of 10.82, was obtained in the simulation reported. This value is superior to the GSM mean compression rate, equal with 8, [12]. This compression rate can be improved if a loss less compression method is added to implements the blocks CO and DE. The proposed method assures a good quality reconstruction (on each segment an output signal to noise ratio superior to 16 dB was obtained and the mean output signal to noise ratio was superior to 19 dB). It can be said that this compression method is transparent. The GSM compression method do not estimate the reconstruction quality. The proposed compression method is also fast, requiring a relative small number of multiplications. For example the computation of the DCPT requires the same multiplications number like the Fast Fourier Transform of the same signal. This is the reason why the proposed compression method can be implemented on a Digital Signal Processor. The system proposed represents a good alternative to the speech compression systems based on the use of the linear prediction.

Acknowledgments

This research was granted by the AUPELF-UREF in a framework entitled Méthodes modernes de traitement du signal pour la compression de données dans les modems haut débit, coordinated by professor Miranda Naformita.

References

[1] C. E. Shannon, “A mathematical theory of communications,” *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[2] A. N. Kolmogorov, “On the Shannon theory of information transmission in the case of continuous signals,” *Trans. IRE*, vol. IT-2, pp. 102–108, 1956.

[3] D. L. Donoho, M. Vetterli, R. A. Devore, and I. Daubechies, “Data compression and harmonic analysis,” *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2435–2476, 1998.

[4] I. Daubechies, “Ten lectures on wavelets,” in *SIAM*, Philadelphia, 1992.

[5] M. V. Wickerhauser, *Adapted Wavelet Analysis. From theory to software*. A. K. Peters Ltd, 1994.

[6] D. Pastor and R. Gay, "Décomposition d'un processus stationnaire du second ordre. Propriétés statistiques d'ordre 2 des coefficients d'ondelettes et localisation fréquentielle des paquets d'ondelettes," *Traitement du signal*, vol. 12, no. 5, pp. 393–420, 1995.

[7] M. Borda and D. Isar, "Whitening with wavelets," in *ECCTD. '97 Conference*, Budapest, 1997.

[8] V. E. Neagoe, "Introducing a new orthogonal spatial transform for significant data selection," *Revue de l'Académie roumaine, Bucharest*, pp. 163–180, 1983.

[9] R. Boite and M. Kunt, *Traitement de la parole*. Lausanne: Presses Polytechniques Romandes, 1987.

[10] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Comm. Pure Appl. Math.*, no. 41, pp. 909–996, 1988.

[11] N. Moreau, *Techniques de compression des signaux*. Paris: Masson, 1995.

[12] ETS 300 580-2, "Full rate speech transcoding (GSM 06.10 version 4.0.2). standarde ETSI," 1994.