2010 First International Conference on Integrated Intelligent Computing

New Strategy of Lossy Text Compression

Shawki A. Al-Dubaee, and Nesar Ahmad
Department of Computer Engineering,
Aligarh Muslim University, Aligarh, India
shawkialdubaee@gmail.com, n.ahmad.ce@amu.ca.in

Abstract— **This paper proposes a new strategy that is based on the signal processing tools applied to text compression of files namely, the wavelet transform and the fourier transform. The influence of compression size and threshold of wavelet filters and the fourier transform as well as two parameters: families of wavelet filters and decomposition levels, on compression factor of text files are investigated. The experimental results are shown that the wavelet and the fourier transforms are suitable for lossy text compression with non-stationary text signal files. In addition, the fourier transform is the most suitable with files which have same characters such as aaa.txt and aaaa.txt files. However, the results of wavelet and fourier transforms are lossless text compression with stationary text signal files (aaa.txt and aaaa.txt files). This research also represents a step forwards dealing with both images and text compression i.e. multimedia compression.**

Keywords- Wavelet transforms; Fourier transforms; ASCII; lossy; Text Compression

I. INTRODUCTION

In recent years, we have witnessed the tremendous growth of textual information via the Internet, digital libraries and archival text data in many applications. Data compression is playing an increasingly significant role in saving storage space efficiently and in accelerating transmission speed. The aim of text compression uses some techniques to represent the digital text data in alternate representations with removing redundant data in order to get less space [1], [2]. Data compression methods are usually classified as either Lossless or lossy methods.

Lossless compression can reconstruct the original data from the compact. Lossless compression algorithms can be classified into three categories [1], [3]: statistical methods (such as Huffman coding [4] and Arithmetic coding [5]), dictionary methods (Lempel/Ziv algorithm [6]) and substitution methods (The Burrows Wheeler Transform (BWT) [7]). Lossy compression generally provides much higher compression than lossless compression but the decoding process only results in a close approximation to the original data. The elimination or replacement of irrelevant characters or words in text considered to be general methods of lossy compression of text. Their applications are archiving of online journals, blogs, instant text messages, emails, speed writings, abbreviations, and one solution of low bandwidth transmission and communication in the computer networks. Witten et al. [8] have suggested using lossy text compression to improve the text compression ratio. They wanted to use one compression method for both image and text compression as one domain. Kaufman and Klein introduced some ideas for lossy text compression [9]. Cannataro et al. [10] developed lossless and lossy compression methods and interpreted XML documents as a datacube using a multidimensional approach. Wavelet transforms can be used as lossy compressor for sequences of numbers in XML documents. Their belief is that lossy compression will play an important part in forthcoming applications on Internet. In [11], Palaniappan and Latifi proposed three techniques for character-based lossy text compression namely, Dropped Vowels (DOV), Letter Mapping (LMP), and Replacement of Characters (ROC). The main field of signal processing is often applied to signals and images [12].There is a real need to find suitable tools to perform indexing and compressing of multimedia at the same time. In addition, the compression of small or very large text files is difficult to obtain [3].

At the same time, one can gain much benefit from adopting lossy text compression using the wavelet transform or the fourier transform. The wavelet and fourier transforms deal with the text files as signals. Forty-two wavelet filters of six families, namely, Haar(haar), Daubechies (db1-10), Biorthogonals (bior1.1-6.8 and rbior1.1-6.8), Cofilets (coif 1-5), Symlets (sym 1-10), and Dmey (dmey) are investigated in order to identify a suitable wavelet function for lossy text compression [23]. Therefore, this is exactly why we decided to present a proposed method and discussion in this paper. The rest of paper is organized as follows. Section II discusses the preliminaries that are related to wavelet transform. In section III, we describe the problem formalization and algorithm. Section IV, shows our database. Section V, contains results and discussions. Finally, section VI provides conclusions and future directions.

II. AN OVERVIEW WAVELET TRANSFORM

Wavelets, the term wavelet translates from ondelltes in French into English which means a small wave [14], are an alternative to solve the shortcomings of Fourier Transform (FT) and Short Time Fourier Transforms (STFT). FT just has frequency resolution and no time resolution which is not suitable to deal with non-stationary and non-periodic signal. In an effort to correct this insufficiency, STFT adapted a single window to represent time and frequency resolutions. However, there are limited precision and the particular window of time is used for all frequencies. Wavelet representation is a lot similar to a musical score that location of the notes tells when and what the frequencies of tones are occurred [13]. This is due to the fact that it has a good time-scale (time-frequency or multi-resolution analysis) localization property having fine frequency and having coarse time resolutions at lower frequency

978-0-7695-4152-5/10 $26.00 © 2010 IEEE
DOI 10.1109/ICIIC.2010.51

IEEE computer society

resolution and coarse frequency and fine time resolutions at higher frequency resolution. Therefore, it makes suitable for indexing, compressing, information retrieval, clustering detection, and multi-resolution analysis of time varying and non-stationary signals.

Wavelets are basis functions that allow transformation of signals from their original domain to another in which some operations can be performed in an easier way. In mathematic, a wavelet is a function $\Psi(t) \in$ L2(R) where L2(R) space is the space of all square-integrable functions defined on the real line R with a basic property [15],[16].

$$\int_{-\infty}^{\infty} \Psi\,(t)\,dt \;=\; 0 \qquad (1)$$

This means that the average value of the wavelet in time domain must be zero, and therefore it must be oscillatory.

In other words, $\Psi(t)$ must be a wave, and the wavelets are generated from a single basic function $\Psi(t)$, called Mother Wavelet, by using both scale a and translate b factors and the constant $|\,a\,|^{-1/2}$ is used for energy normalization [13-14].

$$\Psi_{a,b}(t) = |\,a\,|^{-1/2}\;\Psi\,(\frac{t-b}{a}) \qquad (2)$$

$\Psi_{a,b}(t)$ stands for the wavelet basis.

Wavelet transforms can be divided into the Continuous Wavelet Transform (CWT), the Series Wavelet Transform (SWT) and Discrete Wavelet Transform (DWT) based on the variable **a** and **b** which are continuous values or discrete numbers and type of input signal [15], [16].

A. *Discrete Wavelet Transform (DWT)*

The discrete wavelet transform is also called the series wavelet transform but the difference is that it is used to transform a digital signal as follows [15], [16]:

$$c_{j,k} = \sum_{n=-\infty}^{\infty} x(n)\Phi_{j,k}(n) \qquad (3)$$

$$d_{j,k} = \sum_{n=-\infty}^{\infty} x(n)\Psi_{j,k}(n) \qquad (4)$$

where $c_{j,k}$ and $d_{j,k}$ illustrate both approximation and detail coefficients. Here $\Phi(n)$ is Scaling function

$$\Phi_{j,k}(n) = \sqrt{2^j}\,\Phi(2^j n - k) \qquad (5)$$

Ψ(n) is Mother wavelet function:

$$\Psi_{j,k}(n) = \sqrt{2^j}\,\Psi(2^j n - k) \qquad (6)$$

Furthermore, the original signal can be reconstruction by

$$x(n) = \sum_{k=-\infty}^{\infty} c_{j0,k}\Phi_{j0,k}(n) + \sum_{j=j0}^{\infty}\sum_{k=-\infty}^{\infty} d_{j,k}\Psi_{j,k}(n) \qquad (7)$$

where $j0$ is the last level in the decomposition stage.

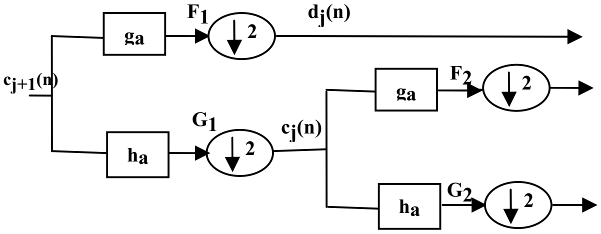
B. *Fast Wavelet Transform (FWT)*

In 1989, Mallat proposed that Multi Resolution Analysis (MRA) can be used to obtain the Discrete Wavelet Transform (DWT) of a discrete signal by replacing the scaling function $\Phi(n)$ and mother wavelet $\Psi(n)$ with low pass and high pass filters respectively. Thus, an increase in speed for the wavelet transform is giving. For increasing of speed the algorithm, Mallat proposed replacement equations (3) and (4) with (8) and (9) respectively [14], [15]:

$$c_j(n) = \sum_{m=-\infty}^{\infty} h_a(m-2n)c_{j+1}(m) \qquad (8)$$

$$d_j(n) = \sum_{m=-\infty}^{\infty} g_a(m-2n)c_{j+1}(m) \qquad (9)$$

Where ha is low pass filter and ga is high pass filter



$$g_a(n) = (-1)^n\, h_a(n-1) \qquad (10)$$

Figure 1. Fig. 1. Two level decomposition of FWT.

As shown in Figure 1, the low pass (h_a) and high pass (g_a) filters achieve convolution of the c_{j+1}(n) input signal and subsequently down sampling them by factor 2. An output of this level is c_j(n) and d_j(n), where c_j(n) is approximation coefficients and d_j(n) is detail coefficients.

III. METHODOLOGIES

The distinctive characteristic of the algorithm is that it converts the original text into signal to build proper compression. As shown in Figure 2, the algorithm consists of six steps:

1- Pre-processing the input text file; this includes reading the original text and converting it into proper format to be ready for processing by one of signal processing tools, fast wavelet transform and fast fourier transform (FFT);

2- Decomposing FWT or FFT and using it to compress the original text file;

3- Applying the range of threshold to obtain an approximation result which is well suited for signal compression;

4- Computing the compression for decomposition of FWT or FFT by sorting, and finding effect of threshold on decomposition, and selecting a small number of insignificant approximation coefficients and changing them into zeroes;

5- Post-processing and measuring the error compression according to original and reconstructed signals; and

6- Repeating the above aforementioned steps to check if there is more than one file.

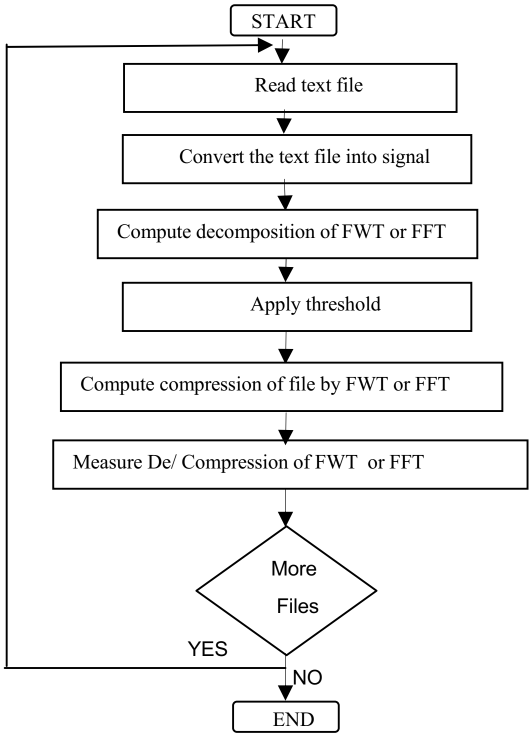

START
Read text file
Convert the text file into signal
Compute decomposition of FWT or FFT
Apply threshold
Compute compression of file by FWT or FFT
Measure De/ Compression of FWT or FFT
More Files
YES
NO
END

Figure 2 Algorithm of proposed text compression method.

Our algorithm reads the text files as ASCII and then it deals with the ASCII files as signals for every text file as shown on figures 3,4,5 and 6 [19], [20], [21].

The FFT has just one filter for decomposition and reconstruction which it applies the algorithm for once of every text file. However, the FWT applies the algorithm for different types of FWT with specified decomposition levels which are used for text compression.

There are forty-two wavelet filters of six families, namely, Haar(haar), Daubechies (db1-10), Biorthogonals (bior1.1-6.8 and rbior1.1-6.8), Cofilets (coif 1-5), Symlets (sym 1-10), and Dmey (dmey) with 5 levels of decomposition that are investigated in order to evaluate a suitable wavelet filter for lossy text compression. The purpose of applying different types of wavelet filters and different levels is to select a suitable filter and level for lossy compression of text.

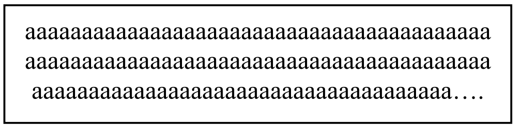

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa….

Figure 3. Fig. 3. Sample of aaa.txt text file of ‘a’ character.

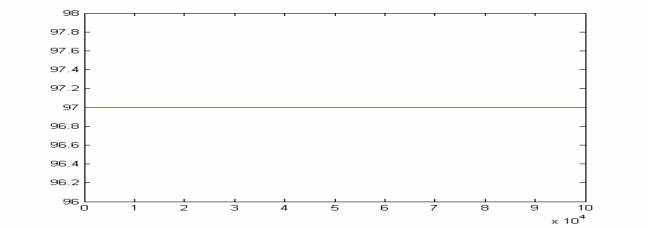


Figure 4. Fig. 4. Stationary signal of ‘a’ character and its ASCII (97) for aaa.txt text file.

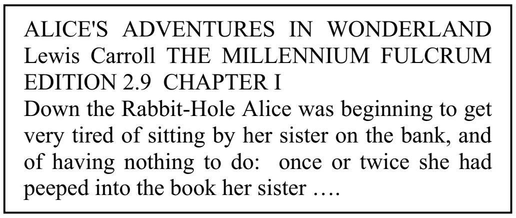

ALICE'S ADVENTURES IN WONDERLAND Lewis Carroll THE MILLENNIUM FULCRUM EDITION 2.9 CHAPTER I
Down the Rabbit-Hole Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister ….

Figure 5. Fig. 5. Sample of alice29.txt text file.

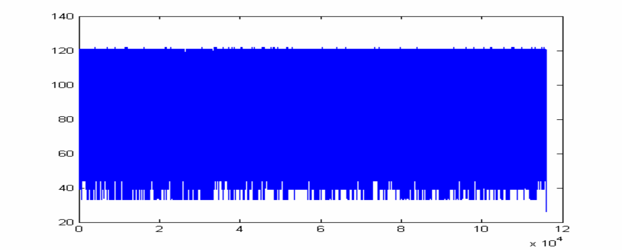


Figure 6. Fig. 6. Non-Stationary signal of alice29.txt text file (text signal file) and its ASCII characters.

Our implementation is evaluated based on compression factor of text files and 2-norm errors of the thresholds (0.6, 0.7, 0.8 and 0.9) are as follows [10]:

$$\text{The compression factor (CF)} = \left(1 - \left(\frac{\text{Compressed Size}}{\text{Original Size}}\right)\right) X100 \qquad (11)$$

$$\text{Error} = \frac{\sqrt{\left(\sum_i |x_i - xc_i|\right)^2}}{\sum_i |x_i|} \qquad (12)$$

Where $x=[x_i \text{ ……… } x_N]$ is the original signal $xc=[xc_i \text{ ……… } xc_N]$ is the reconstructed one. The error rate is between 0 and 1. The threshold determines maximum of a small number of insignificant approximation coefficients in order to change them into zeroes in transformation domain.

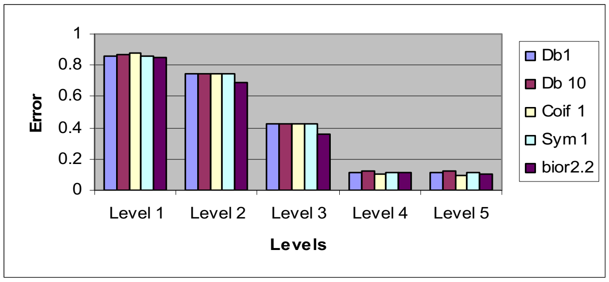


Figure 7. Fig. 7. Error rates and 5 levels for some wavelet filters of bible.txt file for threshold 0.9.

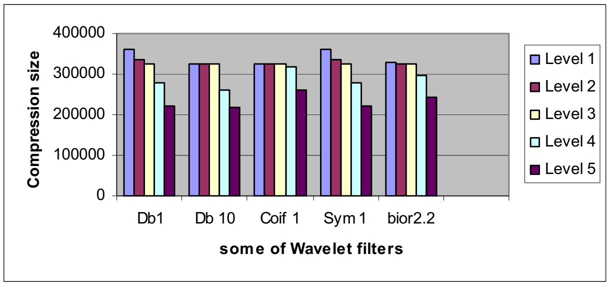


Figure 8. Fig. 8. Effect compression size of some wavelet filters and for 5 levels of bible.txt file threshold 0.9.

IV. Results and Discutions

We apply our experiment on data created from the Calgary/Canterbury text compression corpus to evaluate the practical performance of our method [17], [18]. The result of reading of the five text files and dealing with them as signals are non-stationary signal files except the aaa.txt and aaaa.txt text files which consider as stationary signal files. This is because; the aaa.txt and aaaa.txt text files include same character which is the 'a' character but are at difference sizes, 97.6 KB and 489 KB respectively.

The results of FWT of aaa.txt and aaaa.txt appear that it is not fit to stationary signal files so that the FWT is the most suitable for text files which have different characters which called are non-stationary signal file and FFT is the most suitable for single characters of text files (aaa.txt and aaaa.txt) which are called stationary signal file according to our method as shown on the tables 1, 2 and 3. In addition, the FWT and the FFT compress the stationary signal files (aaa.txt and aaaa.txt) as lossless text compression where the error rate is zero. We can also notice from tables 1, 2 and 3 that the some results of the FFT for some thresholds have been given superior compression than the FWT except the 0.6 threshold of FWT. However, the error rate of the FFT is close to the FWT. We consider that the most suitable of wavelet filter achieves a good error rate regardless of the compression factor whereas the error rate can be improved in future as shown in tables 1, 2 and 3 and figures 7 and 8. Also, the wavelet transform is flexible with selecting the filters than FFT. Therefore, the FFT and the FWT are suitable for lossy text compression with non-stationary signal file.

The compression factor of Coif1 of the threshold 0.6 for all files is more than fourier transform except with paper.txt file whereas they have same CF as shown in table 1. In addition, the CF of fourier transform of threshold 0.9 and alice29.txt and paper1.txt files of threshold 0.6 are higher than wavelet transform. However, the error rates are smaller than wavelet transform.

In figure 8, the compression size of Db1 (Haar), Db10, bibr2.2, and Sym1 filters are better than Coif1, especially the CF of Db10 which has 216137 Values (93%) of bible.txt file for threshold 0.9. However, the error rate of Coif1 is smaller than all filters including the filter of fourier transform so that we prefer to take the coif1 as shown in figure 7.

Therefore, it is expected that wavelet filter of Coif1 can be used for lossy text compression effectively with files that contain different characters which are called non-stationary signal files.

V. Conclusions and Future Directions

After dealing with the text file as signal to facilitate using signal processing. The performance of the wavelet and fourier transforms is tested and the following points are observed. The wavelet and fourier transforms are suitable for non-stationary signals of text compression. However, the fourier transform is the most suitable for stationary signals of text compression. No further advantage can be gained in processing of lossy text compression beyond level 5 for FWT.

As a result, it is expected that wavelet filter of Coif1 can be used for lossy text compression effectively with files that contains different characters which are called non-stationary signal files. Furthermore, the Coif1 wavelet filter will be a good solution to the compression of small or very large text files. The Foruier transform is expected to be suitable for single characters or stationary signal files.

In sum, the proposed method has been improved the error rate of compression and decompression of wavelet transform more than foruier transform. The FWT and FFT compress the stationary signal files (aaa.txt and aaaa.txt) as lossless text compression where the error rate is zero. The wavelet transform is generally used in lossy compression of audio and image data (JPEG 2000) [12], [22]. Moreover, we show the ability of wavelet and fourier transforms to perform lossy text compression so that it can compress all texts, images and audio in Internet application easily and effectively at the same time. Also, using lossy text compression may be one of the solutions of low bandwidth transmission and communication in the computer networks.

For future work, we suggest to enhance and develop our method to perform lossless text compression. We believe that wavelet and fourier transforms will perform well for multimedia compression.

References

[1] K. Sayood, Introduction to Data Compression. 2nd ed., Morgan Kaufmann edition, 2000.

[2] N.Ziviani, E. Moura, G. Navarro, and R. Baeza-Yates, “Compression: a key for next generation text retrieval systems” vol. 33, no.11, pp.37-44 ,2000.

[3] J. Platos, V. Snasel, and E. El-Qawasmeh, “Compression of small text files” Advanced Engineering Informatics. Elsevier, vol.22, pp.410-417 ,2008.

[4] D.A. Huffman, “A method for the construction of minimum-redundancy codes”, In: IRE, pp. 1098—1101,vol. 40 , no.9, 1952.

[5] J.J. Rissanen and G. G. Langdon, “Arithmetic coding, IBM Journal of Research and Development”, vol. 23, no.2, pp.149-162,1979.

[6] J. Ziv and A. Lempel, “Compression of individual sequences via variable rate coding” IEEE Transactions on Information Theory, IT-24, pp.530-536 ,1978.

[7] M. Burrows and D.J. Wheeler, “A block-sorting lossless data compression algorithm”, Technical report. Digital Equipment Corporation, Palo Alto, California ,1994.

[8] I. H. Witten, et al, “Semantic and generative models for lossy text compression.”, The Computer Journal, vol.37 ,no.2 ,1992.

[9] K. Kaufman, and S.T. Klein, “Semi-lossless text compression” International Journal of Foundations of Computer Science, 16(6), pp.1167-1178 ,2005.

[10] M. Cannataro, G. Carelli, A. Pugliese, and D. Sacca, “Semantic lossy compression of XML data”, Internatonal Workshop on Knowledge Representation Meets Databases, in conjunction with International Conference on Very Large Data Bases ,2001.

[11] V. Palaniappan, and S. Latifi, “lossy text compression techniques”, In: the 15 International Workshops on Conceptual Structures, ICCS07, pp.205-210 ,2007.

[12] T. Li, et al., “A Survey on wavelet application in data mining”, SIGKDD Explorations, vol.4 ,no.1, pp.3-32 ,2007.

[13] R. narayanaswami, J. Pang, “Multiresolution analysis as an approach for tool path planning in NC machining”, Elsevier Computer Aided Design, vol. 35, no.2 ,pp.167—178 ,2000.

[14] C. S. Burus, R. A. Gopinath, and H. Guo, “Introduction to wavelets and wavelet transforms”, Prentice Hall Inc., (1998)

[15] S. Mallat, “A theory for multiresolution signal decomposition: the wavelet representation” IEEE Trans. On Pattern Analysis and Machine Intell, vol.11,no.7, pp.674-693 ,1989.

[16] A. graps, “An introduction to wavelets”, IEEE Computational Science and Engineering, vol. 2 no..2, pp.50-61 ,1995.

[17] T.C. Bell, I. H. Witten, and J.G. Cleary, “Text Compression”, Prentice Hall, Englewood Cliffs, NJ, 1990.

[18] R. Arnold, and T. Bell, “A corpus for the evaluation of lossless compression algorithms”, In: J.A. Storer, M. Cohn Eds., IEEE computer society press, Los Alamitos, California, pp. 201-210 ,1997.

[19] S. A. Al-Dubaee, and N. Ahmad, “New Direction of Applied Wavelet Transform in multilingual web information retrieval”, The 5th International Conference on Fuzzy Systems (FSKD’08), IEEE Computer Society Press ,vol. 4, pp.198-202,2008.

[20] S. A. AL-Dubaee, and N. Ahmad, “The Bior 3.1 wavelet transform in multilingual web information retrieval”, The 2008 International Conference on Data Mining (DMIN’08), a track at (WORLDCOMP’08), Las Vegas, Nevada, USA, pp.707-713, 2008.

[21] S. A. Al-Dubaee, V. Snasel, and N. Ahmad, “Wavelet, multiwavelet, and multilingualism on the internet”, The 2009 International Conference on Information and Knowledge Engineering (IKE’09), track at (WORLDCOMP’09) , Las Vegas, Nevada, USA, pp. 716-722, 2009.

[22] A. Skodrqs, C. Christopoulos, and T. Ebrahimi: The JPEG2000 still image compression standard. IEEE Signal Processing Magazine, Los Alamitos, California, pp.201-210 ,1997.

[23] M. Misiti, et al., “Wavelet Toolbox MATLAb user’s guide Version 4”, Mathworks Inc, 2008.

Table 1. Compression Size and Compresson factor of Fourier and Coif 1 filter of wavelet transforms with level 5 l 5 with non-stationary text signal files.

Files	Fourier Transform				Wavelet transform Coif 1 and Level 5			
	0.6	**0.7**	**0.8**	**0.9**	**0.6**	**0.7**	**0.8**	**0.9**
bible.txt 3.85MB (3250898 Values)	648617 (80%)	486211 (85%)	323745 (90%)	161643 (95%)	612832 (81%)	475357 (85%)	359316 (90%)	261510 * (92%)
alice29.txt 149KB (115973 Values)	22915 (80%)	17181 (85%)	11383 (90%)	5659 (95%)	21711 (81%)	17340 (85%)	13609 (88%)	9523 * (92%)
paper1.txt 51.9KB (44309 Values)	8649 (80%)	6469 (85%)	4329 (90%)	2133 (95%)	8699 (80%)	6747 (85%)	5232 (88%)	3411* (92%)
Average	80%	85%	90%	95%	80.7%	85%	88.7%	92%

* Db1 (Haar), Db10, bibr2.2, and Sym1 have differente compression factors.

Table 2. Compression Size and Compresson factor of Fourier and Coif 1 filter of wavelet transforms with level 5 with stationary text signal files.

Files	Fourier Transform				Wavelet transform Coif 1 and Level 5			
	0.6	**0.7**	**0.8**	**0.9**	**0.6**	**0.7**	**0.8**	**0.9**
aaa.txt 97.6KB (100000 Values)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	3129 (97%)	3129 (97%)	3129 (97%)	3129 (97%)
aaaa.txt 489KB (500000 Values)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	15629 (97%)	15629 (97%)	15629 (97%)	15629 (97%)

Table 3. Error rates of Fourier transform and Coif 1 filter of wavelet transforms with Level 5

Files	Fourier Transform				Wavelet transform Coif 1 and Level 5			
	0.6	**0.7**	**0.8**	**0.9**	**0.6**	**0.7**	**0.8**	**0.9**
bible.txt	0.108	0.112	0.118	0.125	0.087	0.089	0.092	0.099
alice29.txt	0.134	0.140	0.147	0.156	0.107	0.109	0.112	0.123
paper1.txt	0.168	0.173	0.180	0.189	0.127	0.129	0.133	0.147
aaa.txt	0	0	0	0	0	0	0	0
aaaa.txt	0	0	0	0	0	0	0	0
Average	0.082	0.0878	0.089	0.094	0.0642	0.0654	0.0674	0.0738