A Simple Feature Reduction Method for the Detection of Long Biological Signals*

Max Chacón[1], Sergio Jara[1], Carlos Defilippi[2],
Ana Maria Madrid[2], and Claudia Defilippi[2]

[1] Departamento de Ingeniería Informática, Universidad de Santiago de Chile,
Av. Ecuador 3659, PO Box 10233, Santiago, Chile
mchacon@diinf.usach.cl, stjara@nt.entel.cl
[2] Hospital Clínico, Universidad de Chile,
Av. Santos Dumont 999, Santiago, Chile
cdefilippi@med.uchile.cl, amadrid@ns.hospital.uchile.cl

Abstract. Recent advances in digital processing of biological signals have made it possible to incorporate more extensive signals, generating a large number of features that must be analyzed to carry out the detection, and thereby acting against the performance of the detection methods. This paper introduces a simple feature reduction method based on correlation that allows the incorporation of very extensive signals to the new biological signal detection algorithms. To test the proposed technique, it was applied to the detection of Functional Dyspepsia (FD) from the EGG signal, which is one of the most extensive signals in clinical medicine. After applying the proposed reduction to the wavelet transform coefficients extracted from the EGG signal, a neuronal network was used as a classifier for the wavelet transform coefficients obtained from the EGG traces. The results of the classifier achieved *78.6%* sensitivity, and *92.9%* specificity for a universe of 56 patients studied.

1 Introduction

The incorporation of more extensive biological signals and of new transformation methods to represent those signals produces a large number of features which are difficult to analyze by the classifying algorithms that allow the detection of a pathology. To overcome these problems there are feature extraction methods such as Principal Component Analysis and Feature Selection by Mutual Information [1-2]. But these methods require a number of cases (at least equivalent to the features to be selected) to carry out the extraction. On the other hand, the incorporation of new pathologies with long signal registers makes it difficult to obtain test subjects for the analyses, decreasing the number of examples. To solve this problem, use of a simple method is proposed that allows a reduction of the number of redundant features according to the degree of correlation existing between them.

To carry out an evaluation, a problem of great clinical interest has been chosen, which also generates a signal made up of very extensive electric registers.

* This work was supported by FONDECYT project Nº 1050082.

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 431 – 439, 2005.

Functional Dyspepsia (FD) is a complex syndrome which can not be detected by clinical examination and affects 25% of the population. At present, the precise nature of the mechanisms that produce this symptomatology is unknown, but it seems unlikely that a single mechanism can explain the variety of discomforts that comprise this syndrome [3].

The lack of knowledge regarding the specific mechanisms that give origin to this syndrome, the necessity of ruling out a variety of alterations, added to the high degree of incidence in the population, highlight the importance of having recourse to efficient diagnostic mechanisms for the detection of FD. The methodology used at present for the identification of FD consists of following the so called Rome protocol based on the systematic elimination of possible organic alterations [4]. This results in costly procedures and long periods during which patients must live with this condition.

A different approach in order to establish minimal motor alterations in these patients came to light about a decade ago, and involves the study of the electric activity of the digestive tract. These studies are based on the analysis of the graphs of electro-gastric activity over time, obtained from surface electrodes placed on the patient's abdomen. The resulting record, which is similar to an electroencephalogram, is called an electrogastrogram (EGG)[4-8].

Spectral analyses carried out by means of a Fourier transformation are the methods most often used for extracting information from electro-gastric activity. The difficulty in recording these signals has resulted in the design of new methods in order to improve the signal/noise ratio of the EGG [3,9]. The long signal records (approximately 2 hours) require block processing which produces undesired averaging effects in the spectra. In order to avoid this problem, special processing techniques have been developed based on adaptive and mobile media models which achieve a significant improvement in the quality of the record [9].

In several papers attempts have been made to evaluate gastric activity by means of an EGG, but these refer to pathologies other than FD, and they focus on the methods of classification (such as the use of neuronal networks), in which the first steps include the use of a classic Fourier analysis or the extraction of parameters from this transform [10-11].

The main disadvantage of analyses based on Fourier transforms for the diagnosis of FD is that they do not have the ability to temporarily locate the phenomenon of interest. This is due to the fact that Fourier theory only possesses frequency ability, and thus, although it is possible to determine the total number of frequencies that make up a signal, it is impossible to determine the time at which they occur. [12]. This problem becomes especially relevant in the study of EGGs related to FD because it is necessary to analyze the gastric system in its different states: *pre-prandial* (before the ingestion of food), *prandial* (during the ingestion of food), and *post-prandial* (after ingestion of food), which results in records that are too long to only analyze frequencies.

In order to solve the problem of time resolution, a variety of solutions have been developed that attempt to provide, to a greater or lesser degree, a simultaneous improvement in time and frequency resolution. Some of these are spectral methods that vary in time, spectro-temporal methods, and time-scale methods. Most of these

solutions are based on segmentation of the signal, thus transforming the problem into a search for the optimal segment.

Among the different alternatives, wavelet transformation stands out because it avoids the problems of segmenting the signal by using windows based on functions that can be completely scaled and modulated. This is called a multiresolution analysis [13]. This type of transform is a powerful alternative for the analysis of non-stationary signals whose spectral characteristics change in time, such as biomedical signals in general [14] and EGG in particular.

Thus, this work consists of pre-processing an EGG signal in order to select the segment that contains FD information, calculating the coefficients of the wavelet transform, and subsequently using them as input for a neuronal classifier which will discriminate between healthy and dyspeptic patients.

2 Methods

2.1 Foundations

In order to avoid the segmentation of the signal required for the Fourier windowing calculation, the wavelet transformation (WT) uses a different alternative that consists of using a window that moves through the signal allowing the spectral calculation of each position. Then we iterate by gently increasing or decreasing the size of the window, thus obtaining a complete set of time-frequency representations at different resolutions.

The WT decomposes the signal into a set of base functions that correspond to a family. Families are generated by dilation and translation of a basic wavelet, called the "mother" wavelet, which is a function of time denoted by decomposes the signal ncy representations at different resolutions..ncreasing or decreasing the size of the window, obtai $\psi(t)$. The translation of ψ provides temporal resolution, and the dilation provides scaling. There are two important conditions that a wavelet must fulfill: i) the function must decay in time $\lim_{t\to\infty}|\psi(t)| = 0$. ii) The function must oscillate so that $\int \psi(t)dt = 0$. In order to implement these functions there are various alternatives, among which the ones most used are those of Haar, Daubechies and Morlet, which are shown in Figure 1.

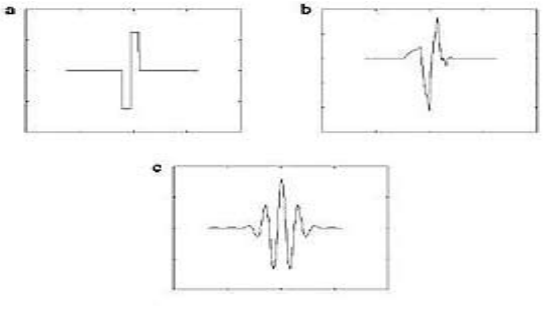

Fig. 1. Wavelets of Haar (a), Daubechies (b) and Morlet (c)

For applications that involve the digital processing of signals, the discrete wavelet transformation (DWT) is used. The result of DWT is a multilevel decomposition in which the coefficients that determine a high degree of resolution correspond to the high frequency components of the signal, while the low resolution levels correspond to the low frequency components.

For the implementation of DWT, beyond the base wavelets that act as bandpass filters, scaling functions, are used to establish upper limits for scaling factors.

The base wavelets in conjunction with the scaling functions form a bank of filters that are applied to the signal to be transformed. The low pass filters formed by the scaling functions limit the spectrum of the base wavelets on the basis of a given scale, covering the lower frequency functions. The output of the filter bank comprises the wavelet coefficient series.

The division of the spectrum is carried out by means of a multiresolution analysis which divides the spectrum in two. The details of the high frequency portion of the signal are kept, while the half corresponding to the lower frequencies can be again subdivided as often as necessary, and is limited only by the available information. Figure 2 below illustrates this type of treatment.

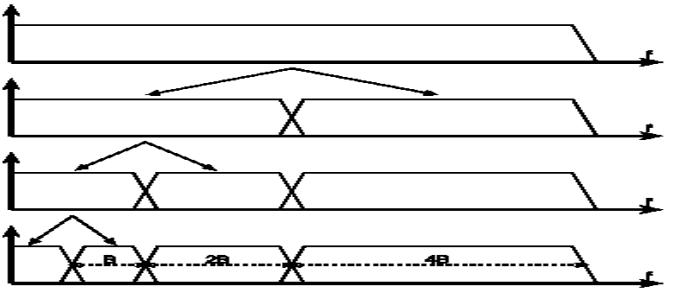

Fig. 2. Division of the spectrum by means of multiresolution analysis

2.2 Data Collection

The original data set corresponds to a total of 150 EGG exams carried out on subjects most of whom suffered from diverse gastric disorders, and among which there is a control group of 14 healthy patients. From the total of sick patients, 42 were selected that fit the Rome protocol; adding the healthy patients to these, a final set of 56 exams is generated for analysis. These exams were carried out between the years 2000 and 2002 in the Clinical Hospital of the Universidad de Chile, using a computational tool known as Polygram Version 5.0 developed by Gastrosoft Inc. [15] for recordring, processing and storing data. The signals obtained were stored digitally with a sampling frequency of 8 Hz for subsequent processing by Matlab version 6.1, using the signal processing, wavelet and neuronal network toolboxes.

Each exam consists of a 2.5-hour record. After a 10 minute relaxation period, the *pre-prandial* stage is initiated under fasting conditions and lasts approximately one hour. Subsequently, a light meal is given to the patient for ingestion, thus initiating

the *prandial* stage which lasts between 20 and 30 minutes. Finally, the *post-prandial* stage begins which lasts approximately one hour.

2.3 Process and Pre-processing

The data obtained from the Polygram equipment presents a very high rate of sampling because the maximum frequencies in the stomach correspond to tachygastric episodes and reach 0.15 Hz or 9 cycles per minute (cpm). Frequencies between 9 and 12 cpm correspond to activity in the small intestine. These signals have frequency components that are outside the range of gastric activity, and include a great deal of noise.

In order to focus the process on the relevant information, a subsampling process is carried out followed by a filtering of the signal. The exam is separated into its three stages (*pre-prandial*, *prandial* and *post-prandial*), in order to calculate the wavelet transform coefficients. The complete process is illustrated in Figure 3.

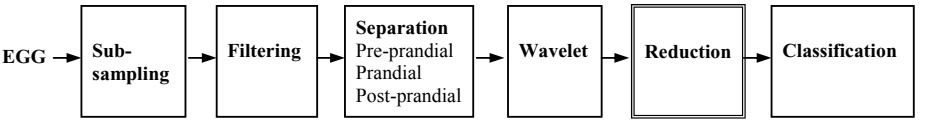

Fig. 3. Pre-processing and classification of EGG

Once the wavelet transform coefficients have been obtained, very little relevant information exists in the low frequency bands (flat responses) and high frequency bands, and these are therefore discarded. After obtaining the coefficients and deleting the high and low frequencies, there still remain a great deal of coefficients for each period which contain redundant information. With this large number of coefficients, and only 56 cases, it is not possible to carry out a Principal Component Analysis or Feature Selection by Mutual Information [1-2].

2.4 Feature Reduction

The reduction method from the generated wavelet coefficients is the following:

i) Create groups of correlated coefficients (given a correlation interval, e.g. 0.05). For that purpose the quantity and kind of wavelet coefficients having a correlation greater than 0.95 are calculated, then those greater than 0.90, and so on successively until a low correlation (e.g. 0.2) is reached. Each calculation begins with the total set of wavelet coefficients. This leads to pairs (c,r), where c represents the number of wavelet coefficients (features) that have a correlation greater than the value indicated by r.

ii) Create a correlation curve, with the correlation index r on the abscissa, with $r \in [0.2, 0.95]$, and the number of coefficients having a correlation coefficient greater than r on the ordinate.

iii) The choice of an adequate point for the reduction will always be a compromise between the reduction of coefficients (features) and the elimination of information. The idea is to try and decrease as much as possible the number of components

without decreasing the information considered in the analysis. More than one point can be chosen and evaluated with the classifier. The points that are candidates to be evaluated will be those that show the largest drop along the curve from right to left.

2.5 Classification

Classification is carried out by means of a static neuronal network which uses the backpropagation method for training. The input layer uses the reduction of the wavelet coefficients, and for the hidden layers zero to two layers are tested. For the output layer two output neurons were evaluated: the implementation of a classic classifier, and an output neuron for which a threshold must be calculated.

Different training methods were evaluated such as backpropagation with momentum, resilient backpropagation, secant and second order methods (Levenberg-Marquardt) [16]. In order to evaluate the training of the network, a cross validation process was used [17] which consisted in separating the initial set into seven groups. Each group consisted of the exams of six dyspeptic and two healthy patients. Training was carried out with six groups, and the seventh was reserved for evaluation. This process was carried out seven times in order to evaluate all groups.

3 Results

3.1 Pre-processing

The process is initiated by subsampling the signal which selects one sample for every 20 original samples, thus obtaining a sampling frequency of 24 cpm.

The signal filtering is carried out with a Butterworth low-pass fifth-order filter with a cutoff frequency of 10 cpm. The purpose of this cut-off frequency is to eliminate small intestine activity (9 cpm to 12 cpm), without damaging the signals that correspond to gastric activity.

The EGG record is divided into the three previously mentioned sections, which are analyzed separately. At this stage it is necessary to ensure that the length of each segment is the same for each patient as a way of normalizing the input to the neuronal classifier.

3.2 Feature Extraction

For the calculation of the DWT, the three base wavelets shown in Figure 1 were tested. Daubechies' wavelet shows the best results. An analysis of variability between subjects shows that low and very high frequency signals do not carry any useful information, and thus these coefficients were discarded as shown in Figure 4 below.

Figure 5 shows the graph of correlation versus coefficients obtained according to section 2.4 for the *prandial* stage. It is seen that there are four points at which the curve has more pronounced drops. These points were evaluated by the neuronal classifier, and the best result was obtained with the point having the coordinates (0.65, 80).

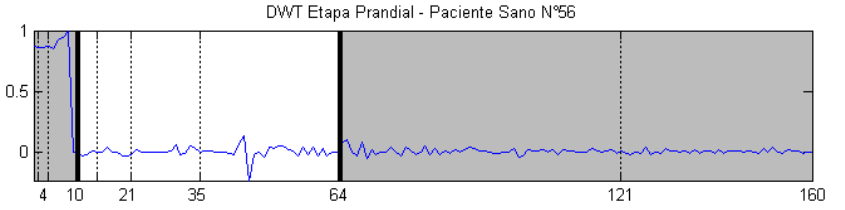

Fig. 4. Elimination of coefficients that do not carry information (shaded regions)

The total coefficient reduction achieved for the *prandial* stage was from 160 to 80, and that for the *pre* and *post-prandial* stages was from 1070 to 300.

3.3 Classifications

Different neuronal classifiers were implemented for each of the exam stages, and the four training methods mentioned in Section 2.5 were evaluated. Sigmoid neurons were used for the hidden layer, and linear and sigmoid neurons were tested for the output layer.

By means of the cross validation process, models with one and two hidden layer were evaluated first, and acceptable results were obtained. However, these results are achieved with reduced numbers of neurons in the hidden layer. Due to this fact, it was decided to eliminate the hidden layer, thus transforming the classifier into a linear discriminator which uses a single output neuron with a sigmoid activation function.

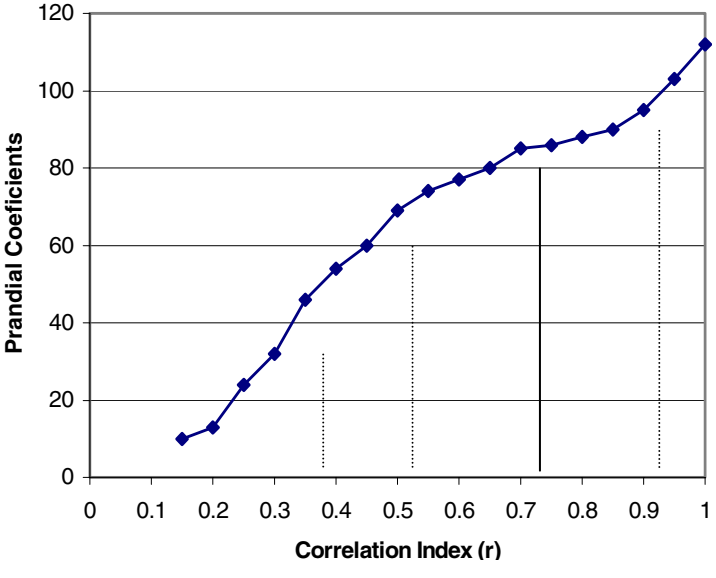

Fig. 5. Feature reduction graph, *prandial* stage (chosen point, solid line)

The best results were obtained for the *prandial* stage with 80 input coefficients, one output neuron, and the resilient backpropagation training method. The threshold value is adjusted in order to improve classification while using only training data, thus achieving *82.1%* accuracy (*17.9%±6%* error, with $p<0.005$), with *78.6%* sensitivity and *92.9%* specificity.

4 Conclusions

Time-frequency analysis methods make it possible to obtain important features for identifying biological signals. When the signals are extensive, however, the number of features generated by these methods prevent an adequate classification of the signals. This paper presents a simple method for extracting important features that make it possible to classify satisfactorily very extensive biological signals. The method has been evaluated using one of the most extensive signals known in clinical practice, that of EGG records (2.5 hours per patient) for the detection of FD.

Attempts to diagnose gastric electrical abnormalities in FD by studying the frequencies generated by the spectral analysis of segments of the EGG signal are not satisfactory. Attempts to systematically extract EGG characteristics for their subsequent classification have generated adequate results in other gastric pathologies [10], but the vast majority of these methods are based purely on a frequency analysis, and complex indices must be developed in order to characterize the different phenomena.

The time-frequency analysis based on the wavelet transform generated more than 1000 coefficients for identifiable sections of EGG signals. Attempts to classify directly these coefficients did not allow an adequate discrimination between healthy patients and those suffering from FD. Only after applying the proposed feature reduction the cases were separated adequately, achieving 82.1% accuracy. In this particular case, application of the feature extraction allowed the complexity of the classification to be reduced to a linear separation problem that was implemented by a neuronal network without hidden layer.

The proposed feature reduction introduced here can be extended to other problems of identification of long biological signals such as those of sleep-wakefulness EEGs [18], or signals of blood pressure and flow for the analysis of the autoregulation of brain blood flow [19].

References

1. Blum A. Langley P. Selection of relevant feature and examples in machine learning, Art. Intell. **97** (1997) 245-71.
2. Duda R. Hart P. Patter Classifications, Wiley, 2nd Eds. (2001).
3. Wai-Man W., Cheng C., Chu-Yu B., Chun-Yu Wong B., Wai-Mo H., Non-Ulcer dyspepsia, Med Progress **2** (2003) 1-8.
4. Drossman D., "ROME II – The Functional Gastrointestinal Disorder". 2nd edition, Degnon Associates, Mc Lean, VA, USA. (2000).
5. Liang J., Chen J., What can be measured from surface electrogastrography (computer simulations). Dig Dis Sci **42** (1997) 1331-43.

6. Verhagen M., Van Schelven L., Samsom M., Smout A., Pitfalls in the analysis of electrogastrographic recording, Gastroenterology **117** (1999) 453-460.
7. Akin A., Sun H., Non-invasive gastric motility monitor: fast electrogastrogram (fEGG). Phys. Measu **23** (2002) 505-519.
8. Chen J., McCallum R., Electrogastrography: Measurement, analysis and prospective application, Med Biol Eng Comput **29** (1991) 339-350.
9. Zhiyue L, Chen J. D. Z, Parolisi S., Shifflett J., Peura D., McCallum R., Prevalence of Gastric Myoelectrical Abnormalities in Patients with Non-Ulcer Dyspepsia and Helicobacter Pylori Infection, Dig Dis Sci **46** (2001) 739-745.
10. Chen, J. Lin, Z. and McCallum, R, A Noninvasive feature-based detection of delayed gastric emptying in humans using neural networks. IEEE Trans Biomed Eng **49** (2000) 409-412.
11. Chen, J. A Computerized data analysis system for electrogastrogram. Comput Biol Med 22 (1992) 45-58.
12. Graps A., An Introduction to Wavelets, IEEE Comput Sci Eng **2** (1995) 2-17.
13. Mallat S., A Theory for Multirresolution Signal Decomposition: The Wavelet Representation, IEEE Trans Patter Anal **11** (1989) 674-93.
14. Torrence Ch. Compo G. A Practical Guide to Wavelet Analysis, Am Meteorol Soc **79** (1998) 61-78.
15. Gastrosoft Inc, Polygram – Software reference manual. Lower GI Edition. USA (1990).
16. Prince J., Euliano N., Lefebre W., Neural and Adaptive System. John Wiley & Sons Inc., New York. (2000).
17. Bishop C. Neural Networks for Pattern Recognition, Oxford Claredon Press. (1995).
18. Flexer A., Gruber G., Dorffner G. A reliable probabilistic sleep stager based on a single EEG signal. Artif Intell Med **33** (2005) 199-207.
19. Panerai R. Assessment of cerebral pressure autoregulation in humans - a review of measurement methods. Physiological Measurement **19** (1998) 305-38.