Initialisation and training procedures for wavelet networks applied to chaotic time series

V. Alarcon-Aquino[1], **O. Starostenko**[1], **J. M. Ramirez-Cortes**[2], **P. Gomez-Gil**[3], **E. S. Garcia-Treviño**[1]

[1]Communications and Signal Processing Research Group, Department of Computing, Electronics, and Mechatronics, Universidad de las Americas Puebla, Sta. Catarina Mártir. Cholula, Puebla. 72820. MEXICO. E-mail: vicente.alarcon@udlap.mx
[2]Department of Electronics Engineering
[3]Department of Computer Science, National Institute for Astrophysics, Optics, and Electronics, Tonantzintla, Puebla, MEXICO

Wavelet networks are a class of neural network that take advantage of good localization properties of multi-resolution analysis and combine them with the approximation abilities of neural networks. This kind of networks uses wavelets as activation functions in the hidden layer and a type of back-propagation algorithm is used for its learning. However, the training procedure used for wavelet networks is based on the idea of continuous differentiable wavelets and some of the most powerful and used wavelets do not satisfy this property. In this paper we report an algorithm for initialising and training wavelet networks applied to the approximation of chaotic time series. The proposed algorithm which has its foundations on correlation analysis of signals allows the use of different types of wavelets, namely, Daubechies, Coiflets, and Symmlets. To show this, comparisons are made for chaotic time series approximation between the proposed approach and the typical wavelet network.

Keywords: Wavelet networks, wavelets, approximation theory, multi-resolution analysis, chaotic time series.

1. INTRODUCTION

Wavelet neural networks are a novel powerful class of neural networks that incorporate the most important advantages of multi-resolution analysis (MRA) [3, 4, and 7]. Several authors have found a link between the wavelet decomposition theory and neural networks (see e.g., [1–3, 5, 6, 9, 13, and 21]). They combine the good localisation properties of wavelets with the approximation abilities of neural networks. This kind of networks uses wavelets as activation functions in the hidden layer and a type of back-propagation algorithm is used for its learning. These networks preserve all the features of common neural networks, like universal approximation properties, but in addition, present an explicit link between the network co-efficients and some appropriate transform.

Recently, several studies have been looking for better ways to design neural networks. For this purpose they have analysed the relationship between neural networks, approximation theory, and functional analysis. In functional analysis any continuous function can be represented as a weighted sum of orthogonal basis functions. Such expansions can be easily represented as neural networks which can be designed for the desired error rate using the properties of orthonormal expansions [5]. Unfortunately, most orthogonal functions are global approximators, and suffer from the disadvantage mentioned above. In order to take full advantage of orthonormality of basis functions, and localised learning, we need a set of basis functions which are local and orthogonal. Wavelets are functions with these features. In wavelet theory we can build simple orthonormal bases with good localisation prop-

erties. Wavelets are a new family of basis functions that combine powerful properties such as orthogonality, compact support, localisation in time and frequency, and fast algorithms. Wavelets have generated a tremendous interest in both theoretical and applied areas, especially over the past few years [3, 4]. Wavelet-based methods have been used for approximation theory, pattern recognition, compression, time-series prediction, numerical analysis, computer science, electrical engineering, and physics (see e.g., [3, 5, 10–13, 20–25]).

Wavelet networks are a class of neural networks that employ wavelets as activation functions. These have been recently investigated as an alternative approach to the traditional neural networks with sigmoidal activation functions. Wavelet networks have attracted great interest, because of their advantages over other networks schemes (see e.g., [6, 9–13, 16, 17, 20–24]). In [20, 22] the authors use wavelet decomposition and separate neural networks to capture the features of the analysed time series. In [13] a wavelet network control is proposed to online learn and cancel repetitive errors in disk drives. In [21, 23] an adaptive wavelet network control is proposed to online structure adjustment and parameter updating applied to a class of nonlinear dynamic systems with a partially known structure. The latter approaches are based on the work of Zhang and Benveniste [1] that introduces a $(1 + 1/2)$ layer neural network based on wavelets. The basic idea is to replace the neurons by more powerful computing units obtained by cascading wavelet transform. The wavelet network learning is performed by the standard back-propagation type algorithm as the traditional feed-forward neural network. It was proven that a wavelet network is a universal approximator that can approximate any functions with arbitrary precision by a linear combination of father and mother wavelets [1, 4].

The main purpose of the work reported in this paper is twofold: first, to modify the wavelet network training and initialisation procedures to allow the use of all types of wavelets; and second, to improve the wavelet network performance working with these two new procedures. The most important difficulty to make this is that typical wavelet networks [1–3, 23] use a gradient descent algorithm for its training. Gradient descent methods require a continuous differentiable wavelet (respect to its dilation and translation parameters) and some of the most powerful and used wavelets are not analytically differentiable [4, 6, 7].

As a result, we have to seek for alternative methods to initialise and to train the network. That is, a method that makes possible to work with different types of wavelets, with different support, differentiable and not differentiable, and orthogonal and non-orthogonal. In the work reported in this paper we propose a new training algorithm based on concepts of direct minimisation techniques, wavelet dilations, and linear combination theory. The proposed initialisation method has its foundations on correlation analysis of signals, and therefore a denser adaptable grid is introduced. The term adaptable is used because the proposed initialisation grid depends upon the effective support of the wavelet. Particularly, we present wavelet networks applied to the approximation of chaotic time series. Function approximation involves estimating the underlying relationship from a given finite input-output data set, and it has been the fundamental problem for a variety of applica-

tions in pattern classification, data mining, signal reconstruction, and system identification [3].

The remainder of this paper is organised as follows. Section 2 briefly reviews wavelet theory. Section 3 describes the typical wavelet network structure. In Section 4, the proposed initialisation and training procedures are reported. In Section 5 comparisons are made and discussed between the typical wavelet network and the proposed approach using chaotic time series. Finally, Section 6 presents the conclusions of this work.

2. REVIEW OF WAVELET THEORY

Wavelet transforms involve representing a general function in terms of simple, fixed building blocks at different scales and positions. These building blocks are generated from a single fixed function called mother wavelet by translation and dilation operations. The continuous wavelet transform considers a family

$$\psi_{a,b}(x) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{x-b}{a}\right) \qquad (1)$$

where $a \in \Re^+, b \in \Re$, with $a \neq 0$, and $\psi(\cdot)$ satisfies the admissibility condition [7]. For discrete wavelets the scale (or dilation) and translation parameters in Eq. (1) are chosen such that at level m the wavelet $a_0^m \psi(a_0^{-m} x)$ is a_0^m times the width of $\psi(x)$. That is, the scale parameter $\{a = a_0^m : m \in Z\}$ and the translation parameter $\{b = kb_0 a_0^m : m, k \in Z\}$. This family of wavelets is thus given by

$$\psi_{m,k}(x) = a_0^{-m/2} \psi(a_0^{-m} x - kb_0) \qquad (2)$$

so the discrete version of wavelet transform is

$$
\begin{aligned}
d_{m,k} &= \langle f(x), \psi_{m,k}(x) \rangle \\
&= a_0^{-m/2} \int_{-\infty}^{+\infty} f(x)\, \psi\left(a_0^{-m} x - kb_0\right) dx \qquad (3)
\end{aligned}
$$

$\langle \cdot, \cdot \rangle$ denotes the L^2-inner product. To recover $f(x)$ from the coefficients $\{d_{m,k}\}$, the following stability condition should exist [4, 7]

$$A \|f(x)\|^2 \leq \sum_{m \in Z} \sum_{k \in Z} \left|\langle f(x), \psi_{m,k}(x) \rangle\right|^2 \leq B \|f(x)\|^2, \tag{4}$$

with $A > 0$ and $B < \infty$ for all signals $f(x)$ in $L^2(\Re)$ denoting the frame bounds. These frame bounds can be computed from a_0, b_0 and $\psi(x)$ [7]. The reconstruction formula is thus given by

$$f(x) \cong \frac{2}{A+B} \sum_{m \in Z} \sum_{k \in Z} \langle f(x), \psi_{m,k}(x) \rangle \psi_{m,k}(x), \tag{5}$$

Note that the closer A and B, the more accurate the reconstruction. When $A = B = 1$, the family of wavelets then forms an orthonormal basis [7]. The mother wavelet function $\psi(x)$, scaling a_0 and translation b_0 parameters are specifically chosen such that $\psi_{m,k}(x)$ constitute orthonormal bases for $L^2(\Re)$ [4, 7]. To form orthonormal bases with good time-frequency localisation properties, the time-scale parameters (b, a) are sampled on a so-called dyadic grid in the time-scale

plane, namely, $a_0 = 2$ and $b_0 = 1$, [4, 7] Thus, substituting these values in Eq. (2), we have a family of orthonormal bases,

$$\psi_{m,k}(x) = 2^{-m/2}\psi(2^{-m}x - k) \tag{6}$$

Using Eq. (3), the orthonormal wavelet transform is thus given by

$$d_{m,k} = \langle f(x), \psi_{m,k}(x) \rangle$$
$$= 2^{-m/2} \int_{-\infty}^{+\infty} f(x)\psi_{m,k}\left(2^{-m}x - k\right)dx \tag{7}$$

and the reconstruction formula is obtained from Eq. (5). A formal approach to constructing orthonormal bases is provided by MRA [4]. The idea of MRA is to write a function $f(x)$ as a limit of successive approximations, each of which is a smoother version of $f(x)$. The successive approximations thus correspond to different resolutions [4].

Since the idea of MRA is to write a signal $f(x)$ as a limit of successive approximations, the differences between two successive smooth approximations at resolution 2^{m-1} and 2 give the detail signal at resolution 2^m. In other words, after choosing an initial resolution L, any signal $f(x) \in L^2(\Re)$ can be expressed as [4, 7],

$$f(x) = \sum_{k \in \mathbb{Z}} c_{L,k}\phi_{L,k}(x) + \sum_{m=L}^{\infty}\sum_{k \in \mathbb{Z}} d_{m,k}\psi_{m,k}(x) \tag{8}$$

where the detail or wavelet coefficients $\{d_{m,k}\}$ are given by Eq. (7), while the approximation or scaling coefficients are defined by

$$c_{L,k} = 2^{-L/2} \int_{-\infty}^{+\infty} f(x)\phi_{L,k}\left(2^{-L}x - k\right)dx \tag{9}$$

Equations (7) and (9) express that a signal $f(x)$ is decomposed in details $\{d_{m,k}\}$ and approximations $\{c_{L,k}\}$ to form a MRA of the signal [4].

3. DESCRIPTION OF WAVELET-NETWORKS

Based on the so-called $(1 + 1/2)$-layer neural network, Zhang and Benveniste introduced the general wavelet network structure [1]. In [2, 23] the authors presented a modified version of this network. The main difference between these approaches is that a parallel lineal term is introduced to help the learning of the linear relation between the input and the output signals. The wavelet network architecture improved in [2] and [23] follows the form shown in Figure 1. The equation that defines the network is given by

$$f(x) = \sum_{i=1}^{N} \omega_i \psi\left[d_i(x - t_i)\right] + cx + \bar{f} \tag{10}$$

where x is the input, f is the output, t's are the bias of each neuron, ψ are the activation functions and finally d's and ω's are the first layer and second layer ($1/2$ layer) coefficients, respectively. It is important to note that $\bar{f}$ is an additional and redundant parameter, introduced to make easier the learning of
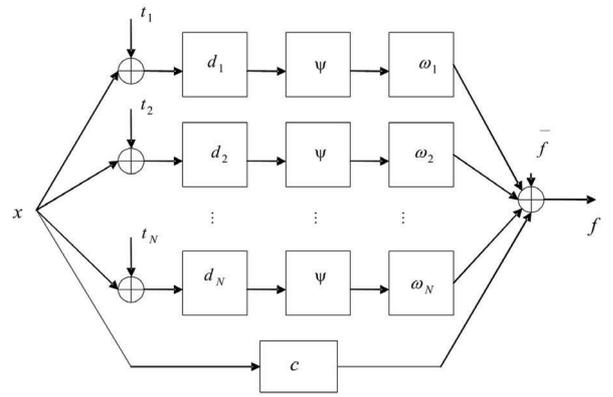
Figure 1 Modified $(1 + 1/2)$-layer wavelet neural network.
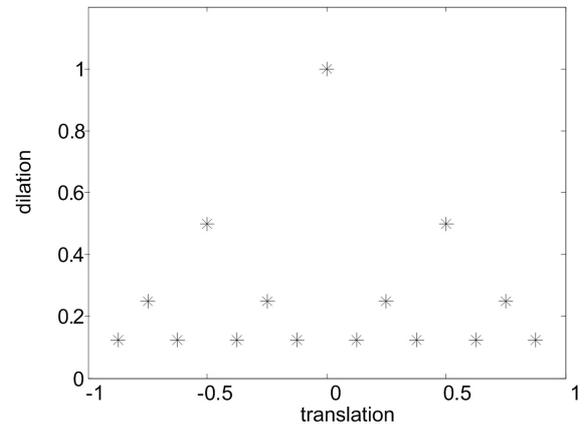
Figure 2 Dyadic Grid for Wavelet Network Initialisation.

nonzero mean functions, since the wavelet $\psi(x)$ is zero mean [4, 7]. Note that Eq. (10) is based on Eq. (1), and therefore the parameters d_i and t_i can be interpreted as the inverted dilation and translation variables, respectively. In this kind of neural network, each neuron is replaced by a wavelet, and the translation and dilation parameters are iteratively adjusted according to the given function to be approximated.

3.1 Wavelet Networks Initialisation

In the works [1, 2 and 23] two different initialisation procedures are proposed. Both based on the idea in some way similar to the wavelet decomposition. Both divide the input domain of the signal following a dyadic grid of the form shown in Figure 2.

This grid has its foundations on the use of the first derivative of the Gaussian wavelet and it is a non-orthogonal grid because the support of wavelet used, at a given dilation, is higher than the translation step at its respective dilation. The main difference between the two initialisation approaches presented in [1, 2 and 23] is the way of the wavelet selection. In the first method, wavelets at higher dilations are chosen until the number of neurons of the network has been reached. In the particular case, that the number of wavelet candidates, at a given dilation, exceeds the remainder neurons, then they are randomly selected using the dyadic grid shown in Figure

2. It is important to note that this method does not take into account the input signal for the selection of wavelets. On the other hand, in the second approach an iterative elimination procedure is used [2, 23]. That is, it is based on the least squares error between the output observations and the output of the network with all wavelets of the initialisation grid. In each iteration, the least contributive wavelet for the least squares error is eliminated of the network until the number of wavelets left in the network is the expected. The number of levels (different dilations) for both cases depends upon the number of wavelets available in the network.

3.2 Training Algorithm

As stated previously, wavelet networks use a stochastic gradient type algorithm to adjust the network parameters. If all the parameters of the network $(c, \bar{f}, d's, t's, \omega's)$ are collected in a vector θ, and using y_k to refer the original output signal and f_θ to refer the output of the network with the vector of parameters θ; the error function (cost function C) to be minimised is thus given by

$$C(\theta) = \frac{1}{2}\mathrm{E}\left\{[f_0(x) - y]^2\right\} \qquad (11)$$

This stochastic gradient algorithm recursively minimise the criterion (11) using input/output observations. This algorithm modifies the vector θ after each measurement in the opposite direction of the gradient of

$$C(\theta, x_k, y_k) = \frac{1}{2}\mathrm{E}\left\{[f_0(x_k) - y_k]^2\right\} \qquad (12)$$

Due to the fact that the traditional procedure used to train wavelet networks is based on the gradient of the error function respect to each parameter of the network, differentiable activation functions are necessary. This is the main reason for the use of the first derivative of the Gaussian wavelet in [1–3 and 23]. Gaussian wavelets are continuous and differentiable wavelets respect to its dilation and translation parameters. In wavelet networks additional processing is needed to avoid divergence or poor convergence. This processing involves constraints that are used after the modification of the parameter vector with the stochastic gradient, to project the parameter vector in the restricted domain of the signal.

4. PROPOSED CORRELATION-BASED INITIALISATION AND TRAINING PROCEDURES

The proposed initialisation procedure is based on the idea of correlation of signals. Correlation is a measure of the degree of interrelationship that exists between two signals [18]. From this definition, wavelet coefficients can be seen as a no normalised linear correlation coefficient, because it is the result of the multiplication and integration of two signals. If the wavelet energy and the signal energy are equal to one, then this coefficient may be interpreted as a normalised linear correlation coefficient. This coefficient represents the similarity between the wavelet and the signal. Higher coefficients

Table 1 Relationship among neurons, decomposition levels, and scales.

Number of Neurons	Decomposition Levels	Scales	Wavelet candidates at a given resolution
1	1	2	2
2	2	2,1	2,3
3	2	2,1	2,3
4	3	2,1,0.5	2,3,5
5	3	2,1,0.5	2,3,5
6	3	2,1,0.5	2,3,5
7	3	2,1,0.5	2,3,5
8	4	2,1,0.5,0.25	2,3,5,9

indicate more similarity [19]. The idea of the proposed initialisation approach consists, firstly, in the generation of the initialisation grid, and secondly, in the selection of wavelets with higher coefficients by using Eq. (7) according to the number of available neurons in the network. The grid is similar to the proposed in [1-3], but in this case a much denser grid (in the translation axis) is used (see Figures 2 and 3). A dense grid is necessary because, at each level of resolution, the whole signal domain can be covered for at least three different wavelet candidates. The number of wavelet candidates at a given resolution depends directly on the effective support of the wavelet. In the work reported in this paper, for comparative purposes and to underline the good performance of this new initialisation technique respect to the traditional wavelet networks training, the first derivative of the Gaussian wavelet is used as well as other types of wavelets. The number of decomposition levels l is defined by the number of neurons of the network as follows:

$$l = int\left(\log_2(N)\right) + 1 \qquad (13)$$

where N is the number of neurons and $int()$ denotes the integer-part function. Note that for the case of a network with one neuron the number of levels is sets to one. The number of wavelet candidates at a given level is then obtained by

$$h_i = \frac{2}{a_i} + 1 \qquad (14)$$

where $a_i = 2^{2-i}$ denotes the scale associated with each level $i = 1, 2, \ldots, l$. Note that equations (13) and (14) can be applied to any wavelet. Regarding the translation parameter of each wavelet candidate, it can be computed as

$$k_{i,j} = (s - n_j)a_i + 1 \qquad (15)$$

where $k_{i,j}$ denotes the translation parameter of the wavelet candidate j at decomposition level i, s represents the wavelet support and $n_j = 0, 1, \ldots, h_i - 1$. For example, for the Daubechies wavelet Db2 with support [0,3], we obtain a grid with 4 levels of decomposition whose scales are 2, 1, 0.5, 0.25 with 2, 3, 5 and 9 wavelet candidates respectively (see Figure 3). Table 1 shows the relationship among the number of neurons, decomposition levels, and the associated scales to each level for a network of eight neurons.
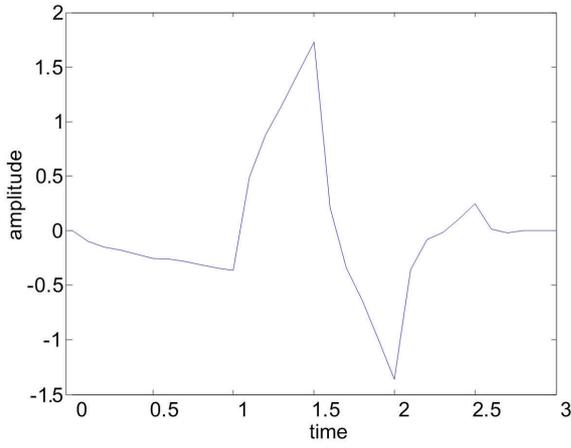
(a)
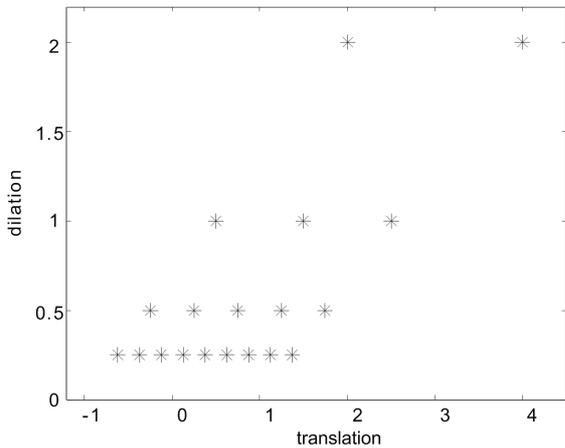
(b)

Figure 3 (a) Daubechies wavelet Db2 and (b) Dyadic grid for wavelet network initialisation for Db2.

Figure 4 shows the initialisation grid for the Gaussian first derivative wavelet. This is a non-orthogonal and dense grid because the support of wavelet used, at a given dilation, is higher than the translation step at its respective dilation. This special characteristic causes the overlap of consecutive translation wavelets. It is worth noting that for all levels the wavelet candidates are distributed in the centre of the signal domain.

4.1 The Training Algorithm

The alternative training approach introduced in this work is based on the concept of direct minimisation techniques. This method works directly with the error function of the network but manipulate only dilation and translation parameters by searching the minimum of the cost function. The rest of the parameters (wavelet coefficients, linear term, and bias) are obtained via linear combination. It is important to note that, similarly to Eq. (10), the equation of the proposed network is based on the continuous wavelet transform, but in this case the parameter d_i corresponds directly to the dilation variable. That is, Eq. (10) can be rewritten as

$$f(x) = \sum_{i=1}^{N} \omega_i \psi \left[\frac{x - t_i}{d_i} \right] + cx + \bar{f} \qquad (16)$$
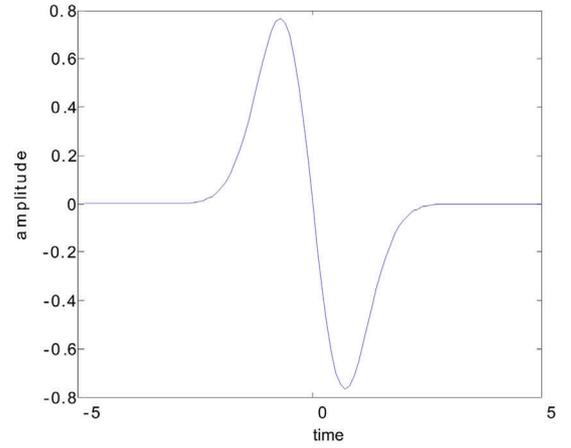
(a)
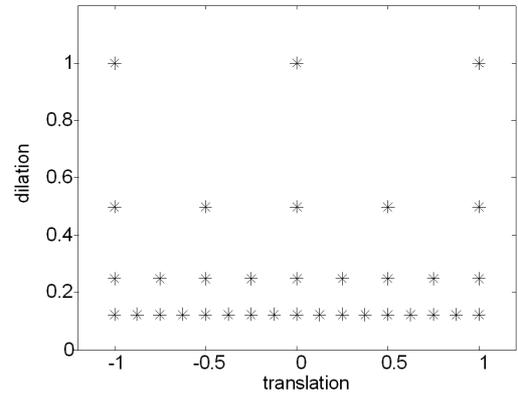
(b)

Figure 4 (a) First derivative of the Gaussian wavelet and (b) Dyadic Grid for Wavelet Network Initialisation for first derivative of the Gaussian wavelet.

If we use $z_i = (x - t_i)/d_i$, the network can be written as

$$f_\theta(x) = \omega_1 \psi(z_1) + \ldots + \omega_N \psi(z_N) + cx + \bar{f}$$

Equation (16) can also be written as

$$f_\theta(x) = W \Psi \qquad (17)$$

where

$$W = \left(\begin{array}{cccccc} \omega_1 & \omega_2 & \cdots & \omega_N & c & \bar{f} \end{array} \right)$$
$$\Psi = \left(\begin{array}{cccccc} \psi_1(z_1) & \psi_2(z_2) & \cdots & \psi_N(z_N) & x & 1 \end{array} \right)^T$$

then finally

$$W = f_\theta(x) \Psi^{-1} \qquad (18)$$

With this strategy the wavelet coefficients, the linear term, and the bias of the network are obtained at the same time. The error function is then given by

$$e = y_k - f_\theta(x_k) \qquad (19)$$

The main idea of this method is the direct modification of the most contributive dilation or translation parameter of the network. Then, indirectly, make the correction of the rest of the parameters via linear combination. For this task, the

special properties of dilation and translation parameters in wavelet theory are taking into consideration. The purpose of this modification is the searching of the highest similarity between the wavelets used to perform the analysis and the signal to be analysed. For higher similarity, a minimum of wavelets can be necessary for the approximation and it will be very precise. As stated previously, this method is based on the idea of direct minimisation techniques because the traditional method used to train typical wavelet networks [1–3 and 23] requires the first derivatives of the error function respect to all its parameters and differentiability is not a property of all families of wavelets [6]. As the traditional optimisation techniques, the reported approach uses a dynamical learning rate which is decreased if the error of the current iteration is poor regarding the error of the last iteration, or increased if the error is higher.

The first step of the process is the generation of all possible combinations of dilation and translation parameters making the modification of only one parameter of them. Since we have N neurons and two parameters, then the number of possible combinations is $2N$. Searching in both directions, positive and negative, then we have $4N$ combinations. Writing a dilation and translation squares matrices with length defined by the number of neurons of the network N we have,

$$
D_{i,i} = \begin{bmatrix} d_1 & d_1 & \dots & d_1 \\ d_2 & d_2 & \dots & d_2 \\ \vdots & \vdots & \ddots & \vdots \\ d_N & d_N & \dots & d_N \end{bmatrix},
$$

$$
T_{i,i} = \begin{bmatrix} t_1 & t_1 & \dots & t_1 \\ t_2 & t_2 & \dots & t_2 \\ \vdots & \vdots & \ddots & \vdots \\ t_N & t_N & \dots & t_N \end{bmatrix} \tag{20}
$$

and using the dilation matrix and the learning rate r, the square matrix of changes can be written as

$$
\Delta_{i,i} = r\left(diag\left(D_{i,i}\right)\right) = \begin{bmatrix} rd_1 & 0 & \dots & 0 \\ 0 & rd_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & rd_N \end{bmatrix} \tag{21}
$$

Concatenating the matrices of Eqs. (20) and (21) to form new matrices of $N \times M$ where $M = 4N$ and consequently $j = 1, 2, \dots, M$, we have

$$
D_{\Delta i,j} = \begin{bmatrix} D_{i,i} + \Delta_{i,i} & D_{i,i} - \Delta_{i,i} & D_{i,i} & D_{i,i} \end{bmatrix}
$$

$$
T_{\Delta i,j} = \begin{bmatrix} T_{i,i} & T_{i,i} & T_{i,i} + \Delta_{i,i} & T_{i,i} - \Delta_{i,i} \end{bmatrix} \tag{22}
$$

Using the matrices of Eq. (22) the different outputs of the network for every combination of dilation and translation parameters are obtained,

$$
f_{\theta j}(x) = \sum_{i=1}^{N} \omega_i \psi\left[\frac{\left(x - T_{\Delta i,j}\right)}{D_{\Delta i,j}}\right] + cx + \bar{f} \tag{23}
$$

where $\omega_i, t_i, d_i, \bar{f}, c \in \Re; i \in Z$. The second step of the process is the evaluation of the error described by Eq. (19) for every combination of Eq. (23). As a result, an error vector is obtained

$$
\begin{aligned}
E_j &= y_k - f_{\theta j}(x_k) = E_1, E_2, \dots, E_{4N} \\
&= [y_k - f_{\theta 1}(x_k), \dots, y_k - f_{\theta 4N}(x_k)]
\end{aligned} \tag{24}
$$

where x_k and y_k are the network input and the desired output respectively, and $f_{\theta j}$ is the output of the wavelet neural network. Finally, searching for the minimum error of this group of errors the most contributive dilation or translation parameter is obtained. It worth noting that the modification of both dilation and translation parameter exclusively depends on the dilation. This is due to the fact that the dilation parameter has a direct relationship with the wavelet support.

5. PERFORMANCE EVALUATION

5.1 Chaotic Time Series

Chaos is the mathematical term for the behaviour of a system that is inherently unpredictable. Unpredictable phenomena are readily apparent in all areas of life [14]. Many systems in the natural world are now known to exhibit chaos or non-linear behaviour, the complexity of which is so great that they were previously considered random. The unravelling of these systems has been aided by the discovery, mostly in this century, of mathematical expressions that exhibit similar tendencies [15]. One might argue that the many factors that influence this kind of systems are the reason for this unpredictability. But chaos can occur in systems that have few degrees of freedom as well. The critical ingredient in many chaotic systems is what mathematicians call sensitive dependence to initial conditions. If one makes even slightest change in the initial configuration of the system, the resulting behaviour may be dramatically different [14].

Chaos is part of an even grander subject known as dynamics. Whenever dynamical chaos is found, it is accompanied by nonlinearity. Naturally, an uncountable variety of non-linear relations is possible, depending perhaps on a multitude of parameters. These non-linear relations are frequently encountered in the form of difference equations, mappings, differential equations, partial differential equations, integral equations, or even sometimes combinations of these. We note that, for each differential equation, the specific parameters were selected because they are the more representative values for chaotic behaviour, and also are the most commonly used in the literature [8, 14, and 15]. In this section, we describe briefly the chaotic time series used in this work. They were considered because they are used as a benchmark and also it illustrates how complex behaviour can easily be produced by simple equations with non-linear elements and feedback.

Hénon Attractor

The Hénon Attractor was introduced by M. Hénon in 1976, and it was derived from a study of chaotic functions trajectories. The Hénon equation is defined by

$$
\frac{dx}{dt} = a + by - x^2, \qquad \frac{dy}{dt} = x \tag{25}
$$

where $a = 1.4, b = 0.3$. Recall that these specific parameters values were selected because they are the more representative for chaotic behaviour, and also are the most commonly used in the literature [8, 14, 15].

Lorenz Attractor

The Lorenz attractor was introduced by E. N. Lorenz, in 1963, and it was derived from a simplified model of atmospheric interactions. The system is most commonly expressed as three coupled non-linear differential equations,

$$\frac{dx}{dt} = a(y - x), \qquad \frac{dy}{dt} = x(b - z) - y, \qquad \frac{dz}{dt} = xy - cz \tag{26}$$

where $a = 10, b = 28, c = 8/3$.

Mackey-Glass

It is the nonlinear a time delay differential equation described by

$$\frac{dx}{dt} = \frac{ax(t - T)}{1 + x^c(t - T)} - bx \tag{27}$$

where $a = 0.2, b = 0.1, c = 10$. Depending on the values of the parameters, this equation displays a range of periodic and chaotic dynamics.

Rössler Attractor

It was introduced by Otto Rossler in 1976, but the originally theoretical equations were later found to be useful in modelling equilibrium in chemical reactions. The system is described by three non-linear ordinary differential equations

$$\frac{dx}{dt} = -y - z, \qquad \frac{dy}{dt} = x + ay, \qquad \frac{dz}{dt} = b + z(x - c) \tag{28}$$

where $a = 0.15, b = 0.20, c = 10$.

Logistic Map

It was explored by ecologists and biologists who used to model the dynamics of the population. The map was popularized in a seminal 1976 paper by the biologist Robert May. Mathematically, the logistic map is written

$$x_{n+1} = ax_n(1 - x_n) \tag{29}$$

where $a = 4$, and the initial condition $x(0) = 0.000104$

5.2 Simulation Results

We want to state firstly that for this work and for testing purposes the wavelet network reported in [1–3 and 23] is defined as wavenet I, whilst the wavelet network with the correlation-based initialisation and the most contributive parameter training procedure proposed in this work is defined as wavenet II. Particularly, the work reported in this paper presents wavelet-networks applied to the approximation of chaotic time series. For this purpose comparisons are made between wavenet I and wavenet II. Comparisons with the back-propagation network can be found in [16]. The wavenet I was tested with

(a)

(b)
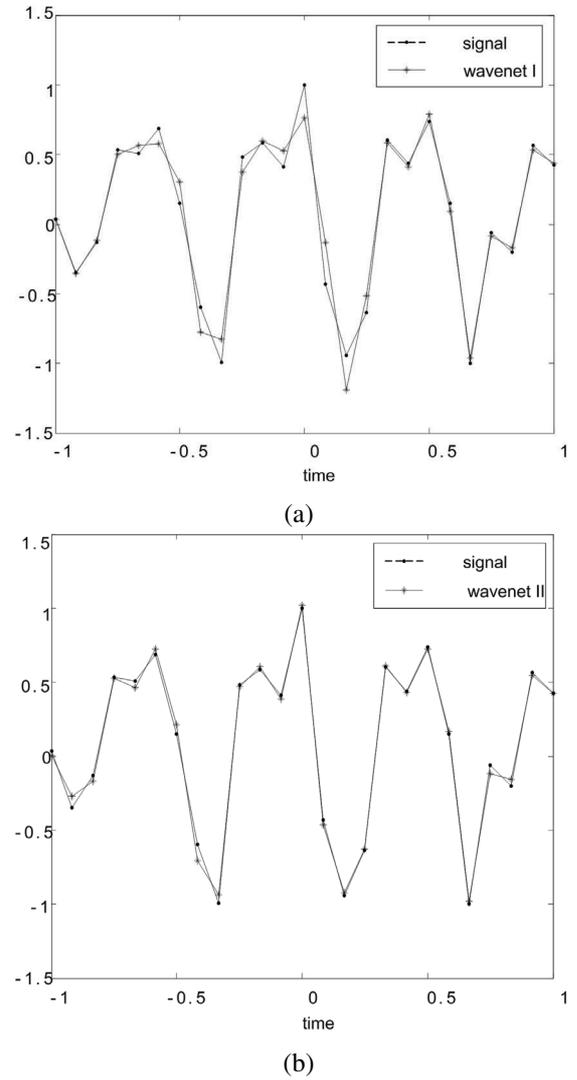
Figure 5 Lorenz Attractor Approximation with 25 samples, 10 neurons and 10 iterations: (a) wavenet I, (b) wavenet II.

the first derivative of the Gaussian wavelet (Gauss1). On the other hand, the wavenet II was tested with the first and second derivative of the Gaussian wavelet (denoted Gauss1 and Gauss2 respectively) as well as Daubechies wavelet 2 (Db2), Coiflets 2 (Coif2), and Symmlet 4 (Sym4). This shows that the proposed approach makes possible to work with different types of wavelets, with different support, not differentiable, and orthogonal and non-orthogonal. The wavelet networks (wavenet I and wavenet II) were assessed with ten neurons and ten iterations. These wavelet networks were also assessed with more number of neurons; however, similar results were obtained (see e.g., [16]).

The Henón and the Logistic functions were sampled every 1 time steps, while the Lorenz equation every 0.01 time steps and sub-sampled every 30 steps. The Rössler attractor was sampled every 0.01 time steps and sub-sampled every 150 steps. The Mackey-Glass equation was sampled every 1 time steps and sub-sampled every 10 steps with a sample period T set to 18. For an appropriate learning, and to take into account the features of the activation functions used by the networks analysed, the series were normalised in the range of $[-1, 1]$ in both time and amplitude domain. As can be seen in Table

Table 2 Simulation Results.

Time Series	Approach	Wavelet	MSE		
			25 samples	50 samples	100 samples
Hénon Attractor	wavenet I	Gauss1	0.1789	0.1140	0.2621
	wavenet II	Gauss1	0.0621	0.1859	0.2248
		Gauss2	0.0878	0.1642	0.2692
		Db2	0.1009	0.1651	0.2696
		Coif2	0.0977	0.1676	0.2790
		Sym4	0.0975	0.1666	0.2777
Lorenz Attractor	wavenet I	Gauss1	0.0319	0.0710	0.0935
	wavenet II	Gauss1	0.0145	0.0672	0.1116
		Gauss2	0.0217	0.0733	0.1073
		Db2	0.0241	0.0660	0.0983
		Coif2	0.0210	0.0788	0.0900
		Sym4	0.0198	0.0843	0.0954
Mackey Glass	wavenet I	Gauss1	0.0315	0.0787	0.1922
	wavenet II	Gauss1	0.0172	0.1809	0.2173
		Gauss2	0.0177	0.1058	0.2278
		Db2	0.0119	0.1101	0.2107
		Coif2	0.0154	0.1290	0.2052
		Sym4	0.0168	0.0973	0.2046
Logistic Map	wavenet I	Gauss1	0.1136	0.1411	0.3188
	wavenet II	Gauss1	0.0669	0.2533	0.3445
		Gauss2	0.0683	0.2377	0.3333
		Db2	0.0734	0.2686	0.3503
		Coif2	0.0156	0.2695	0.3715
		Sym4	0.0217	0.2521	0.3527
Rossler Attractor	wavenet I	Gauss1	0.0090	0.0763	0.2102
	wavenet II	Gauss1	0.0486	0.1974	0.1111
		Gauss2	0.0380	0.1935	0.1419
		Db2	0.0283	0.1980	0.2114
		Coif2	0.0234	0.2155	0.2150
		Sym4	0.0174	0.1762	0.2260

2 and Figure 5, the wavenet II outperforms the wavenet I in terms of MSE (Mean Square Error) for the case of 25 samples in Hénon and Lorenz attractor chaotic time series. This is due to the fact that correlation-based selection of wavelets in a dense grid allows initialising the network with a more correlated wavelets respect to the input signal. For the case of 50 samples, the wavenet I outperforms the wavenet II for the Hénon attractor and Mackey-Glass time series, and for the case of 100 samples the wavenet II shows a better performance than the wavenet I when using the Coiflet wavelet in the Lorenz attractor time series. In summary, we can say that both approaches show somehow similar performance for the approximation of chaotic time series. Note that in some cases wavenet I shows better performance (Mackey-Glass, Rössler and Logistic Map) than the wavenet II (Hénon and Lorenz). Nevertheless, the initialisation and training procedures introduced in this paper show the flexibility of the network for choosing different types of wavelets.

It is well-known from wavelet theory that if there is a major similarity between the wavelet used for the analysis and the signal to be analysed, then a better analysis is performed by the wavelet decomposition. It is important to underline that this can be realised with a minimum of wavelets. In the particular case of the signal and the wavelet used for this work,

we use the translation and dilation properties to generate different versions of the wavelet, and then search for the higher similarity among them and the signal, but working with the corresponding wavelet coefficients. Note that the proposed training algorithm takes into consideration the rule of wavelets selection described previously, and then modify only one dilation or translation parameter of the network searching the minimum of the error function. Furthermore, this modification has the purpose to increase the similarity between the signal and a given most contributive wavelet. This criterion guarantees the convergence of the learning process.

6. CONCLUSIONS

In this paper we have presented initialisation and training procedures for wavelet networks. The results reported in this paper show clearly that the proposed wavelet network has better approximation properties than its similar typical wavelet networks. The reason for this is the firm theoretical link between multi-resolution analysis and correlation theory. The understanding of this interrelation can be used to form a basis which has the capability to explicitly represent the behaviour of a function a different resolutions of input variables with a

minimum of wavelets. An additional advantage of the initialisation and training procedures introduced in this paper is the flexibility of the network to use different types of wavelets. Complex wavelets that include magnitude and phase information may also be used as neurons in the hidden layer of wavelet networks [25]. From a general point of view, wavelet networks can be used for black-box identification of general non-linear systems. These were inspired by both neural networks and wavelet decomposition. The basic idea is to replace the neurons by more powerful computing units obtained by cascading an affine transforms. However, as were presented in this work, wavelet networks can improve its well-known properties with the consideration of the important role played by the correlation theory in wavelet analysis. Finally, it is worth noting that for a comparable number of neurons, the complexity of the input/output mapping realised by the wavelet network reported in this paper is higher than its counterpart realised by typical wavelet networks.

Acknowledgements

REFERENCES

1. Q. Zhang, A. Benveniste, Wavelet Networks, *IEEE Transactions on Neural Networks*. Vol. 3, No. 6, July 1992.

2. Q. Zhang, Wavelet Network: The Radial Structure and an Efficient Initialisation Procedure, *In European Control Conference (ECC 93)*, Groningen, The Netherlands, 1993.

3. S. Sitharama Iyengar, E.C. Cho, Vir V. Phoha, *Foundations of Wavelet Networks and Applications*, Chapman & Hall/CRC. U.S.A. 2002.

4. S. G. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Third Edition, Academic Press, 2008.

5. V. Alarcon-Aquino, E. S. García Treviño, R. Rosas-Romero, J. F. Ramírez-Cruz, L. G. Guerrero-Ojeda, and J. Rodriguez-Asomoza, Wavelet-Network Based on the L_1-norm minimisation for Learning Chaotic Time Series, *Journal of Applied Research and Technology*, Vol. 3, No. 3, December 2005.

6. S. H. Ling, H. H. C. Iu, F. H. F. Leung, and K. Y. Chan, Improved Hybrid Particle Swarm Optimized Wavelet Neural Network for Modeling the Development of Fluid Dispensing for Electronic Packing, *IEEE Transactions on Industrial Electronics*, Vol. 55, No. 9, September 2008.

7. I. Daubechies, *Ten Lectures on Wavelets*, New York. SIAM. 1992.

8. S. H. Strogatz, *Nonlinear Dynamics and Chaos*, Addison Wesley Publishing Company, USA, 1994.

9. E. A. Rying, Griff L. Bilbro, and Jye-Chyi Lu, Focused Local Learning With Wavelet Neural Networks, *IEEE Transactions on Neural Networks*, Vol. 13, No. 2, March 2002.

10. X. Gao, F. Xiao, J. Zhang, and C. Cao, Short-term Prediction of Chaotic Time Series by Wavelet Networks, *WCICA 2004, Fifth World Congress on Intelligent Control And Automation*, 2004.

11. L. Deqiang, S. Zelin, and H. Shabai, A Wavelet Network Based Classifier, *In Proceedings of the 7th IEEE International Conference on Signal Processing ICSP 04*, September 2004.

12. M. Yeginer, Y. P. Kahya, Modeling of Pulmonary Crackles Using Wavelet Networks, *In Proceedings of the 27th IEEE Engineering in Medicine and Biology Conference*, Shanghai, China, September, 2005.

13. C. M. Chang and T. S. Liu, A Wavelet Network Control Method for Disk Drives, *IEEE Transactions on Control Systems Technology*, Vol. 14, No. 1, January 2006.

14. R. L. Devaney, *Chaotic Explosions in Simple Dynamical Systems, the Ubiquity of Chaos*, Edited by Saul Krasner. American Association for the Advancement of Science. Washington DC., U.S.A. 1990.

15. J. Pritchard, *The Chaos CookBook: A Practical Programming Guide*, Part of Reed International Books. Oxford. Great Britain. 1992.

16. E. S. Garcia-Trevino, V. Alarcon-Aquino, and J. F. Ramirez-Cruz, Improving Wavelet-Networks Performance with a New Correlation-based Initialisation Method and Training Algorithm, *In Proceedings of the 15th IEEE International Conference on Computing, CIC '06*. November 2006.

17. E. S. Garcia-Trevino, V. Alarcon-Aquino, Single-Step Prediction of Chaotic Time Series Using Wavelet-Networks, *In Proceedings of the IEEE Electronics, Robotics and Automotive Mechanics Conference, CERMA 06*, September 2006.

18. D. K. Lindner, *Introduction to Signal and Systems*, McGrawHill International Edition, Electrical Engineering Series, USA 1999.

19. A. V. Oppenheim, A. S. Willisky, *Signals and Systems*, 2nd Edition, Prentice Hall, 1996.

20. V. Alarcon-Aquino, J. A. Barria, Multiresolution FIR Neural Network Based Learning Algorithm Applied to Network Traffic Prediction, *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Review*, Vol. 36, Issue No. 2, March 2006.

21. J.-X. Xu, Y. Tan, Nonlinear Adaptive Wavelet Control Using Constructive Wavelet Networks, *IEEE Transactions on Neural Networks*, Vol. 18, No. 1, January 2007.

22. Y. Chen; Luh, P.B.; C. Guan; Y. Zhao; Michel, L.D.; Coolbeth, M.A.; Friedland, P.B.; Rourke, S.J.; Short-Term Load Forecasting: Similar Day-Based Wavelet Neural Networks, *IEEE Transactions on Power Systems*, Vol. 25, No. 1, February 2010.

23. Cheng-Jian, Lin, Nonlinear Systems Control Using Self-Constructing Wavelet Networks, *Applied Soft Computing*, Vol. 9, No. 1, January 2009.

24. R. N. Mahanty and P. B. Dutta Gupta, ANN based Fault Classifier with Wavelet MRA Generated Inputs, *International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications*, Vol. 16, No. 2, June 2008.

25. V. Alarcon-Aquino, O. Starostenko, J. M. Ramirez-Cortes, R. Rosas-Romero, J. Rodriguez-Asomoza, O. J. Paz-Luna, and K. Vazquez-Muñoz, Detection of Microcalcifications in Digital Mammograms Using the Dual-Tree Complex Wavelet Transform, *International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications,* Vol. 17, No.1, March 2009.