Available online at www.sciencedirect.com
ScienceDirect
Journal of Sound and Vibration 302 (2007) 981–999
JOURNAL OF SOUND AND VIBRATION
www.elsevier.com/locate/jsvi

Fault diagnosis of rotating machinery based on auto-associative neural networks and wavelet transforms

Javier Sanz[a], Ricardo Perera[b,*], Consuelo Huerta[b]

[a]CITEAN, Tajonar 20, 31006 Pamplona, Navarra, Spain
[b]Department of Structural Mechanics, Technical University, José Gutiérrez Abascal 2, 28006 Madrid, Spain

Received 15 June 2006; received in revised form 10 January 2007; accepted 11 January 2007
Available online 20 February 2007

Abstract

This paper presents a new technique for monitoring the condition of rotating machinery from vibration analyses. The proposed method combines the capability of wavelet transform (WT) to treat transient signals with the ability of auto-associative neural networks to extract features of data sets in an unsupervised mode. Trained and configured networks with WT coefficients of nonfaulty signals are used as a method to detect the novelties or anomalies of faulty signals. The effectiveness of the proposed technique is evaluated using the numerical data and experimental vibration data of a gearbox. Despite the fact that noise is present in both cases, results demonstrated that the proposed method is a good candidate to be used as an online diagnosis tool for rotating machinery.

1. Introduction

Rotating machines play an important role in industrial applications. Typical applications are in aeronautical, naval and automotive industries. The need to decrease the downtime on production machinery and to increase reliability against possible failures has attracted interest in the online condition monitoring of these systems in recent years. The main purpose of the diagnosis is to analyse the relevant external information in order to judge the condition of the inaccessible internal components so as to decide if the machine needs to be dismantled or not. Although acoustic signal analysis is quite common for the detection of faults in geared systems, vibration-based diagnosis has been more widely used. According to this technique, the presence of a fault will be indicated by changes in the vibration signals picked up from the gearbox casing.

Although sometimes the fault appears to be clearly reflected in a machine's vibration signal, its characteristic features are usually hidden in the vibration signal and, therefore, a sensitive technique for fault signature is needed. Most techniques have sought to represent the machine vibration signal in either the time domain or the frequency domain. The time synchronous average (TSA) providing an average time signal of one individual gear over a large number of cycles [1] has been acknowledged as a powerful and very successful tool in the detection of gear faults [2,3] since it can remove the background noise and all the periodic events

*Corresponding author. Tel.: +34 913363278; fax: +34 913363004.
E-mail address: perera@etsii.upm.es (R. Perera).

0022-460X/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsv.2007.01.006

that are not exactly synchronous with the gear of interest. The use of the residual signal obtained by removing the regular gear meshing harmonics from the TSA is also being frequently used as a fault diagnosis technique [4,5]. The resulting residual signal contains essentially the portion that is caused by the gear fault and geometrical irregularity.

A time–frequency analysis offers an alternative method to signal analysis by presenting information in the time–frequency domain simultaneously. The method known as short-time Fourier transform (STFT) and proposed by Gabor [6] is probably the most widely used time–frequency representation. The characteristic feature of the STFT is the application of the Fourier transform to a time varying signal when the signal is viewed through a narrow window centred at a time t. In this way, the frequency content is obtained at time t and in any other time if the process is repeated. The resolution depends on the size of the window, and as it is constant, a high resolution in time and frequency cannot be obtained simultaneously. So, the window must be chosen for locating sharp peaks or low frequency features and, therefore, its resolution is often unsatisfactory.

For this reason, a more flexible method is required. Wavelet transform (WT) uses more general functions than the sinusoid functions of the Fourier transform as the basis on which a signal is constructed. Originally developed at the end of the 1980s [7–9], WTs are well known for their capability to treat transient signals and have been generating increasing interest in recent years as a tool for fault detection both in machinery [5,10] and in civil engineering structures [11–15] and in other engineering fields [16]. The advantage of wavelet analysis, as opposed to Fourier analysis, is that a WT decomposes a signal into a series of short duration waves or local basis functions (wavelets) on the time axis [17–19], which allows the analysis of local phenomena in vibration signals, such as those caused by faults.

Although visual inspection of certain features of WT can be suitable for identifying damage in some particular situations, for a reliable and automated detection it is necessary to develop a more suitable diagnosis tool. Fault diagnosis is essentially a kind of pattern recognition, or classification. Artificial neural networks (ANN) are a valuable pattern-recognition method in theory and in application. Because of this, models based on neural networks have been applied in recent years in the detection and diagnosis of rotating machinery [4,20–23]. Neural networks can be trained on measured response signals of healthy and damaged specimens to recognise the actual condition of the structure.

The neural network architecture will depend on which level of identification is required. To detect the occurrence of damage, a neural network based on the novelty detection technique can be used [24,25].

The objective is to monitor a sequence of patterns for a healthy structure under normal conditions. These patterns are used as both input and output to train the network. If a signal differs significantly from the herd, then the occurrence of this novelty or anomaly means that damage is alarmed. This approach can be performed by using an auto-associative neural network (AANN).

This paper presents a robust fault detection method in rotating machinery. The proposed method allows an online fault diagnosis and furthermore, unlike other techniques, such as those based on TSA, it works under different operational conditions, a different angular speed and torque transmitted. This is why the signal amplitude wavelet map, in conjunction with an auto-associative neural network, has been used to assess the condition of the machine. An approach for gear fault detection, combining wavelet map patterns with supervised multilayer neural networks, was proposed in Ref. [26]. The advantage of using an unsupervised NN like AANN is that the different patterns used during the training stage do not need to be known previously as this task is performed automatically. Furthermore, the method described in this paper provides a vector of novelty indexes unlike other procedures based on a unique parameter, and this vector shows us which wavelet coefficients are affected by a perturbation in our rotating machinery; therefore, as we are monitoring the time–frequency domain, we can offer either warnings or localisations of the perturbations, and in this sense we can provide a level 2 diagnostic according to Rytter's classification [27].

The layout of the paper is as follows. In Section 2, we will first give some highlights of how the AANNs can provide a measure of novelty or novelty index. In the following sections, we will describe concisely the wavelet theory and discuss how it will be used in this work as a methodology for detecting damage. The effectiveness of the proposed method will be investigated through a numerical simulation study and an experimental study performed on a pump rotor. At the end of the paper we will summarise the present study with conclusions and suggestions for future work.

2. Auto-associative neural networks

Novelty or anomaly detection can be used as a philosophy for damage detection purposes in machines or structures. According to this philosophy, if a new pattern of measured data in the machine or structure differs from previously measured patterns under normal conditions a clear symptom of damage appears i.e., novelty will indicate the presence of a fault.

Among the different methods of novelty detection, the approach based on the use of AANNs is taken here [25]. These kinds of networks correspond to a form of feed-forward multilayer perceptron (MLP) networks configured to reproduce at the output layer those patterns that are present at the input, i.e., the targets used to train the network are simply the input vector themselves, so that the network is attempting to map each input vector onto itself. It represents a form of unsupervised training, as no independent target data is provided.

This network is designed with a bottleneck hidden layer (Fig. 1), i.e. with fewer nodes than the input and output layers, with the purpose of enforcing the network to learn or capture the most significant features or principal components of the input patterns by eliminating their redundancies. This type of configuration with only one hidden layer provides a linear mapping in the bottleneck layer. If, besides the bottleneck layer, two nonlinear hidden layers are also included between the input layer and the bottleneck layer (encoding layer) and between the bottleneck layer and the output layer (decoding layer) (Fig. 1), respectively, nonlinear mapping is found [28,29].

The training of the AANN will involve finding the values of the connection weights as well as the optimum number of neurons in the encoding and decoding layers which minimise an error function between the actual network output and the corresponding input values in the training set. During the training of the network only measured data of the healthy structure are used. The most commonly used training algorithms are based on back-propagation as error function, a bias-variance trade-off coming from the decomposition of the error into bias and variance components [28,30], has been used.

2.1. Novelty index vector

Once the AANN has been trained, the training pattern vector $\mathbf{y}_i$ is fed again into the network yielding an output pattern vector $\mathbf{y}_o$. The residual vector obtained by measuring the difference between the input and an output vector is known as the novelty index vector:

$$\boldsymbol{\lambda} = |\mathbf{y}_o - \mathbf{y}_i|. \tag{1}$$

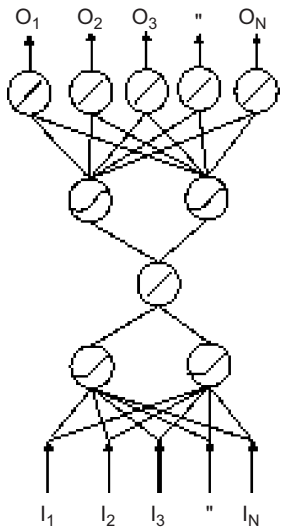

Fig. 1. Nonlinear auto-associative neural network.

The procedure is repeated for the entire sequence of training data vectors obtaining a mean value vector $\boldsymbol{\lambda}_{\text{mean}}$ and a standard deviation vector $\boldsymbol{\sigma}$ for each component of the novelty index vector. From these values, a threshold to judge if damage occurs can be defined as follows:

$$\boldsymbol{\delta} = \boldsymbol{\lambda}_{\text{mean}} + 4\boldsymbol{\sigma}. \tag{2}$$

During the testing stage, new sequences of measured data for the same machine (undamaged and damaged) are fed into the trained network. For each input data vector, a novelty index vector can be defined as previously done. If this index is higher than or equal to the threshold vector $\boldsymbol{\delta}$, then the measured data were taken from a damaged machine.

3. Signal pre-processing using WTs

Faults in rotating machinery originate small perturbations in the vibration signal collected by the transducers. Moreover, different kind of faults are associated with different bands of frequency and, therefore, a suitable procedure which guarantees a good sensitivity to local-global events is necessary as a diagnosis tool. WTs provide a powerful tool for showing local features of a signal and will be dealt within this paper.

3.1. The continuous WT

The continuous wavelet transform (CWT) of a signal $f(t)$ is defined as a convolution integral of $f(t)$ with scaled and dilated versions of a basic wavelet function, called the 'mother wavelet' $\psi(t)$:

$$C_{ab} = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi\left(\frac{t-b}{a}\right) \mathrm{d}t, \tag{3}$$

where a and b are the dilation and translation parameters, respectively. Sharp transitions in $f(t)$ will create wavelet coefficients C_{ab} with large amplitudes, which will be used as the basis to detect faults.

On the other hand, the original signal can be recovered using the inverse CWT as follows:

$$f(t) = \frac{1}{K_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} C_{ab}\psi_{ab}(t)\, \mathrm{d}b \frac{\mathrm{d}a}{a^2}, \tag{4}$$

where K_ψ is a constant dependent on the wavelet type.

3.2. The discrete WT

The CWT is highly redundant, which means that not all the coefficients are necessary to reconstruct the original signal. This redundancy requires a significant amount of computation time and resources. A more efficient and compact form of wavelet analysis, with a significant reduction in computation time, can be reached by decomposing the signal into a discrete subset of translated and dilated parent wavelets, usually taking as dilation and translation parameters integer powers of two ($a = 2^j$, $b = k\ 2^j$), which corresponds to dyadic sampling. Furthermore, this kind of sampling, unlike other kinds of discrete wavelets, is shift invariance like the CWT, which means that the WTs of a signal and of a dyadic scale or time-shifted version of the same signal are simply shifted versions of each other.

Using the discrete scales, the discrete wavelet transform (DWT) is defined as follows:

$$C_{jk} = 2^{-j/2} \int_{-\infty}^{\infty} f(t)\psi(2^{-j}t - k)\, \mathrm{d}t = \int_{-\infty}^{\infty} f(t)\psi_{jk}(t)\, \mathrm{d}t. \tag{5}$$

In an analogous way, with the CWT the original signal can be rebuilt using the inverse discrete wavelet transform (IDWT):

$$f(t) = \sum_{j \in Z} \sum_{k \in Z} C_{jk}\psi_{jk}(t). \tag{6}$$

If in Eq. (6) it is considered that the WT C_{jk} is only available up to the level J, the original function can be written as

$$f(t) = \sum_{k=-\infty}^{\infty} C_{Jk}^{A} \phi_{Jk}(t) + \sum_{j \leqslant J} \left(\sum_{k=-\infty}^{\infty} C_{jk} \psi_{jk}(t) \right) = A_J + \sum_{j \leqslant J} D_j, \tag{7}$$

where ϕ_{Jk} is the scaling function at level J, A_J is called the approximation function or sub-signal at level J, $D_j(t)$ is detail sub-signal defined for each level j and C_{Jk}^{A} are the level-J approximation coefficients obtained as follows:

$$C_{Jk}^{A} = \int_{-\infty}^{\infty} f(t) \phi_{Jk}(t)\, \mathrm{d}t. \tag{8}$$

In Eq. (7), the wavelets cover the spectrum up to scale J, while the rest is done by the scaling function. Therefore, the number of wavelets is now limited. This type of decomposition is called multiresolution analysis as it generates a hierarchical set of approximation and detail sub-signals, which give information about the trends and fluctuations, respectively, at different scales (frequencies).

As for most functions the WTs do not have analytical solutions, and they are calculated numerically by means of the fast wavelet transform (FWT) developed by Mallat [8]. Since, as was commented above, the WT can be seen as a filter bank, this algorithm uses a series of high and low pass filters to progressively find the wavelet and scaling function transform coefficients. More details about this algorithm can be found in Strang and Nguyen [31].

For the study performed in this paper, the detail sub-signals are extremely relevant as they are most sensitive to the changes or fluctuations of the vibration signal originated by faults. Furthermore, the suitable scales will be those corresponding to integer multiples of the fault frequency (harmonics). In this way, we will consider those scales smaller than the scale related to the meshing frequency of the gear which will allow all the frequencies higher than the meshing frequency to be checked, particularly the harmonics of the tooth frequency.

DWT has some properties or characteristic features which make it specially suitable for our purposes of fault detection. Among others, the most relevant are the localisation and the characterisation of the local regularity of the signal subjected to analysis [19,32,33].

4. AANN based wavelet methodology for online fault diagnosis of rotating machinery

The principle of the proposed method is feature extraction by DWT of the collected vibration signal and then, from these features, identification of significant changes or novelty detection using an auto-associative network. As it is assumed that damage will alter the measured patterns, novelty index vector will indicate a fault and its localisation (angular shaft position where the fault is located).

Therefore, the methodology proposed for pattern-recognition-based online fault diagnosis is performed in three modules or stages. The first module is the measurement and signal pre-processing using DWT; the second module corresponds to the training of the AANN and, finally, the last stage is the recognition module using the previously trained AANN (Fig. 2).

The resulting diagnosis system will have the following characteristics:

1. The signals are measured from the casing of the rotating machine.
2. Automatic and online diagnosis.
3. Identification of faults under different operational conditions, different angular speed and torque transmitted.
4. Partial localisation of the damaged machine component, i.e., if the nature of the failing component produces different frequency events that are not multiples of the frequencies generated by another potential failing part, then we will be able to determine which rotating element is deteriorating. In this way, for

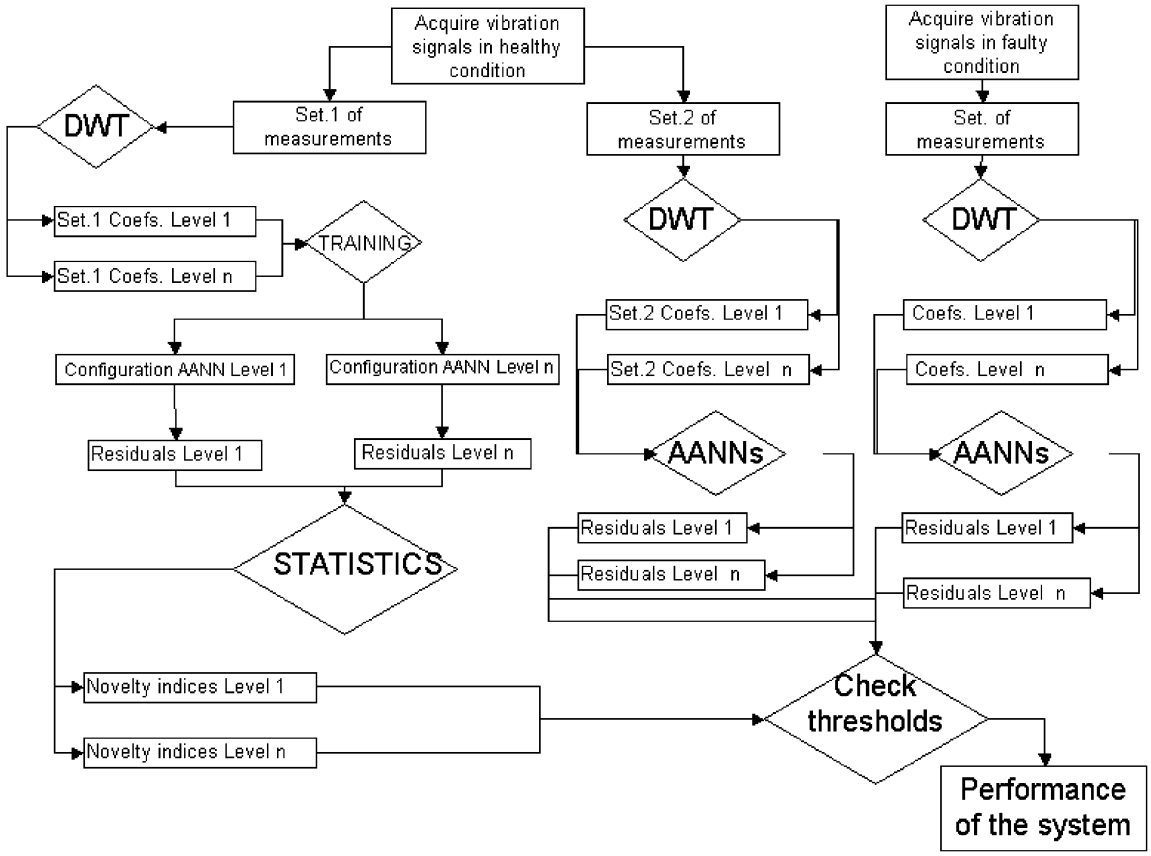

Fig. 2. Flowchart of the proposed diagnosis method.

instance, if a gearbox has different gears, the identification of the damaged one will be successful if its number of teeth does not coincide or is not a multiple of the number of teeth of the other gears.

5. Robustness, since the background noise does not affect the diagnosis (if the level or amplitude of this contribution is lower than the amplitudes caused by pulses originated by irregular operational conditions).

4.1. Measurement and signal pre-processing

By application of WT, vibration signals in time domain can be transformed to time–frequency signals. The main idea behind the use of wavelets is based on the fact that the presence of local faults introduces short-duration changes in the vibration signal, which can be observed from the distribution of the wavelet coefficients for the CWT or the detail signals for the DWT. The procedure to follow during this first stage is the following:

1. Measure the vibration signal of the machine during every revolution of the shaft. The transducer's location must be appropriate for monitoring the faults being investigated.
2. The second step requires the selection of the most suitable wavelet for analysis and its level of decomposition. Since it is difficult to characterise a pulse caused by abnormal conditions, we do not have any information for choosing the most optimal wavelet. The selection is usually done by trial and error. However, we know that differences among various faults correspond to different levels of frequencies and, therefore, this can be used to establish the level at which the wavelet analysis must be performed in the case of the DWT. Therefore, different levels of decomposition can be defined for different kinds of faults.

3. From the measured signal, compute the detail signal coefficients of the DWT according to the level of decomposition chosen previously. These coefficients carry, in a compact way, information of the main local and global features of the signal and are therefore, taken as the input feature vectors to the classifying network.

4.2. Training of the AANN

The implementation of the auto-associative neural network for damage detection requires two modules or stages: The training stage and the detection stage. Vibration data of intact machinery obtained analytically or experimentally, are used as training samples to train the AANN. During this stage, the weights are adjusted and the configuration of the artificial neural network is defined; then the novelty index vectors are determined. The procedure is as follows:

1. One AANN is defined for each level of decomposition adopted in the previous stage. The detail coefficient vectors of the DWT are taken as input and output parameters of the network. In order to adjust the weights and the number of neurons in the encoding and the decoding layers for each AANN, for the given training sets, a minimisation of an error based on a natural trade-off between bias and variance is used [30].
2. Definition of the novelty index vector. To do this, once each network has been trained, the training pattern vectors are again introduced to the AANN. Then a residual vector is defined from the difference between each one of the output and input vectors. With all the residual vectors determined in this way, threshold parameters or novelty index vectors are determined statistically for each component of these vectors.

4.3. Recognition

The trained network is able to predict faults when a new set of measured data is presented as input to the trained network. To this end, in this stage different vibration signals, corresponding to normal and anomalous operational conditions of the machine, are used as input to the trained networks to check their sensitivity and robustness as a fault detection tool. The procedure is as follows:

1. Analysis of the different vibration signals using the DWT in order to obtain the detail coefficients.
2. Introduction of the detail coefficient vectors to the trained AANN and determination of the residual vectors by subtraction between the output and input vectors.
3. Estimation of the existence or non-existence of machine fault by comparing the residual vectors with the threshold parameters which were calculated in the training stage. The testing results will give us an idea of the effectiveness of the proposed method as an online fault diagnosis tool for our rotating machinery.

5. Case studies

The ability of the proposed method to detect faults is evaluated with two different cases. The first data set is numerically simulated while the second is entirely experimental. In both cases the methodology proposed in the previous section is used.

The simulated case is based on assumed signals under three different operational conditions of a nonlinear system; the introduction of distortions in some of these signals allow evaluation of the performance of the diagnosis method proposed in this paper. In the second case, this methodology is applied to a real rotating machine; this is a gear box in a pump station, where the pitting damage is studied.

In both cases, the robustness of the method is evaluated by checking either the rate of false warnings or the rate of success for damaged signals at different operating points.

For all the steps involved in the diagnosis procedure, i.e., signal pre-processing, training and recognition, specialised software was developed in MATLAB.

5.1. Example 1: numerical study

For the numerical study, three dissimilar signals representative of three different measurements of a rotating machine, recorded at every revolution of the shaft, were taken. Each one of these signals, s1, s2 and s3, is characterised by its amplitude (10, 35 and 50) and frequency (10, 7 and 1 Hz), respectively. In the time domain, the distribution is represented by one rotation of the pinion, which included 51 signal samples every 0.5 Angular Units (Fig. 3).

In order to study the basic functionality of damage recognition, 600 healthy data sets were obtained by making copies of the signals corresponding to normal condition and distorting each copy with different noise vectors since, in reality, measured signals will be degraded by various sources of noise. In the absence of any prescription, Gaussian noise with standard deviation equal to one and zero mean was added to the uncorrupted vibration signals at every sample point. From the 600 data sets, 400 were used for training and validation of the AANN parameters, and the remaining 200 patterns were presented to the AANN in order to check the false warnings of the system.

The fault in the signal has been simulated by adding a perturbation to the raw signals using a cubic spline distributed in different shaft locations. Furthermore, perturbations of different duration (angular supports between 0.03 and 0.05 Angular Units) and amplitude (values between 1 and 12) have been considered, finally generating a set of 1404 perturbed signals, which will be used to evaluate the ability of the AANN to detect faults (Fig. 4).

5.1.1. Signal pre-processing

In this stage the coefficients of the DWT must be calculated with the purpose of their being used as input feature vectors of the AANN. To do this, the vibration signals with or without perturbations were transferred to a computer where a MATLAB program was used to transform each signal into the wavelet domain. To apply the WT, it is necessary to previously select the most suitable family of wavelets as well as the level of decomposition required. To perform this, although there is no defined criterion and the selection is usually done by trial and error, it is necessary to take into account the characteristics of the vibration signals and also the perturbation when it is known.

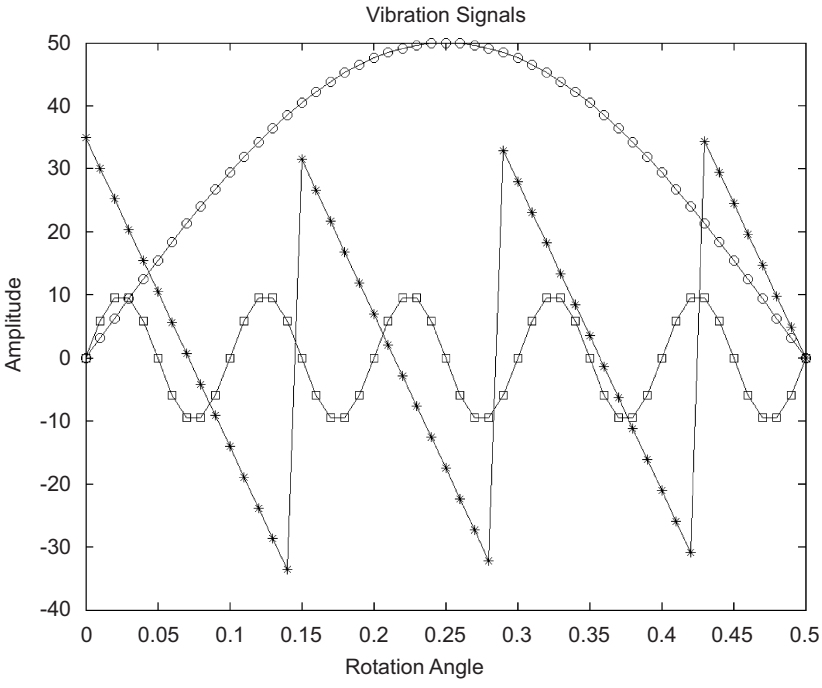

Fig. 3. Uncorrupted vibration signals: '□' s_1 signal, '*' s_2 signal, '○' s_3 signal.

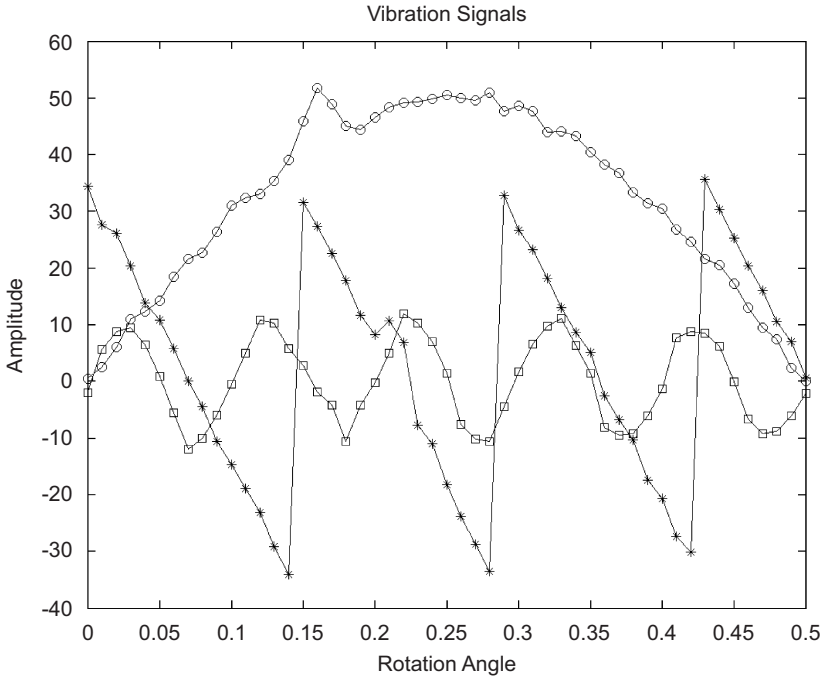

Fig. 4. Perturbations introduced on the corrupted signals: '□' s_1 signal, '*' s_2 signal, '○' s_3 signal.

Assuming as possible only those wavelets that allow an FWT to be carried out, the evaluation of their most relevant properties should be taken into account to guide the wavelet selection. Among these properties, regularity, the number of vanishing moments and support were considered.

According to the first property, Haar wavelets were discarded because of their irregularity as the introduction of artificial discontinuities on the original signal, which would decrease the robustness of the detection system, would not be desirable.

The consideration of the number of vanishing moments of wavelets is very important too because it determines the order of the polynomials that can be approximated. Wavelets with $k+1$ vanishing moments produce zero or very close to zero wavelet coefficients for polynomial signals of k order, i.e., it ensures the suppression of signals that are polynomials of angular units lower or equal to k. According to this, we are interested in the families of wavelets providing small coefficients when the nonperturbed original signals are transformed; however, when perturbations are added, the coefficient values should be high enough to reflect these changes. In the problem subjected to study, the original signals can be approximated using linear or quadratic polynomials, and the perturbations are superimposed using cubic splines. Therefore, the candidate wavelets should have three vanishing moments with the purpose of showing a high value changes originated by the perturbations.

The support of a function is defined as the smallest space-set (or time-set) outside of which function is identically zero. This feature allows identification of those coefficients, which are affected by an event located in a particular time or frequency. For our purpose, we are interested in wavelets having the widest support in frequency and, therefore, the most compact support in time set, i.e., in the angular domain, since it allows a more reliable estimation of the location of the perturbations.

According to the previous three criteria, biorthogonal wavelets of order 3.9 were chosen for the analyses, where 3 and 9 are the wavelet orders for reconstruction and modification, respectively. These wavelets constitute a set of compactly supported biorthogonal spline wavelets with three vanishing moments.

Once the wavelets have been selected, it is necessary to establish the suitable number of levels of decomposition which will be based on the nature of the signal. As the perturbations have been included on the original signals using angular supports between 0.03 and 0.05, the corresponding frequencies of interest will be included in the 20–60 Hz range. According to this range, the wavelet coefficients are obtained from

Eq. (5) for wavelet scales 2, 3, 4 and 5 [33], which will result in 27, 23, 21 and 20 coefficients, respectively, for each level.

5.1.2. Training

After the feature extraction using DWT, the next step is the training of the AANNs. According to the four levels of decomposition of the raw signal performed with DWT, four different damage recognition networks will be established, one for each level.

As reported in Section 2, the AANN configuration has the same number of input and output nodes, which correspond for each one of the four networks to the number of wavelet coefficients obtained for each level of decomposition. Therefore, in order to configure the topology of each one of the four AANNs, only the appropriate number of neurons in the encoding and decoding layers has to be determined. In general, it is not straightforward to determine the best size of the networks for a given system. It may be found only through a process of trial and error. This is performed by configuring different NN architectures with a different number of neurons, finally choosing those configurations which turn out to be more optimal. To this end, starting with a random weight of each connection the resilient propagation learning algorithm is used for the network training and validation in order to adjust the weight functions of the connections by minimising an iterative procedure in which we increase the number of neurons, and for every possible architecture, we choose the optimum by minimising the bias-variance error function between the output and input vectors.

To determine the optimal number of neurons in the encoding and decoding layers, configurations from one neuron to a number approximately equal to 70 per cent of the number of input neurons for each AANN have been considered.

Finally, the following AANN topologies have been taken for each level of wavelet decomposition:

- Level 2: $27 \times 5 \times 1 \times 5 \times 27$
- Level 3: $23 \times 9 \times 1 \times 9 \times 23$
- Level 4: $21 \times 10 \times 1 \times 10 \times 21$
- Level 5: $20 \times 10 \times 1 \times 10 \times 20$

One of the advantages of this kind of NNs is their ability to extract the main features of all the training patterns, which allows the reduction of the noise injected in the training patterns. This can be observed in Figs. 5–8 in which the wavelet coefficients corresponding to the four levels of decomposition of all the training pattern vectors are shown as input (Figs. 5a–8a) and output (Figs. 5b–8b) of the AANNS. It is evident, comparing the inputs with the corresponding outputs, that the AANNs are able to extract the nonlinear features of the three signals, which represents an advantage for detecting any perturbation due to damage.

Once the four networks have been trained, it is necessary to establish the threshold parameters, i.e., the novelty indexes. This is performed for each input coefficient to the resulting NN; therefore, a vector of novelty indexes for each level of decomposition of wavelets is obtained. To do this, the training pattern vectors are introduced again to the trained AANNs, and a set of residual vectors from the difference between each one of the output and input vectors is obtained. The threshold values for each coefficient are calculated from these vectors according to Eq. (2).

5.1.3. Recognition testing

In this stage, the trained AANNs were tested using 200 healthy patterns and 1404 perturbed patterns. The main purpose of this step is to evaluate the ability of these NNs to detect perturbations in the signals produced by faults.

To this end, firstly, we input the 200 healthy testing signals corrupted by noise that have not been used in the training stage, with the purpose of detecting wrong predictions or false warnings in the network outputs. Once these signals have been pre-processed with four levels of decomposition, the wavelet coefficients are inputted in the trained AANNs, and the rate of false warnings given is shown in Table 1. From the results, it can be said that this rate is very low, especially for the three highest levels, which ensures the reliability of the NNs in the case of non-perturbed signals.

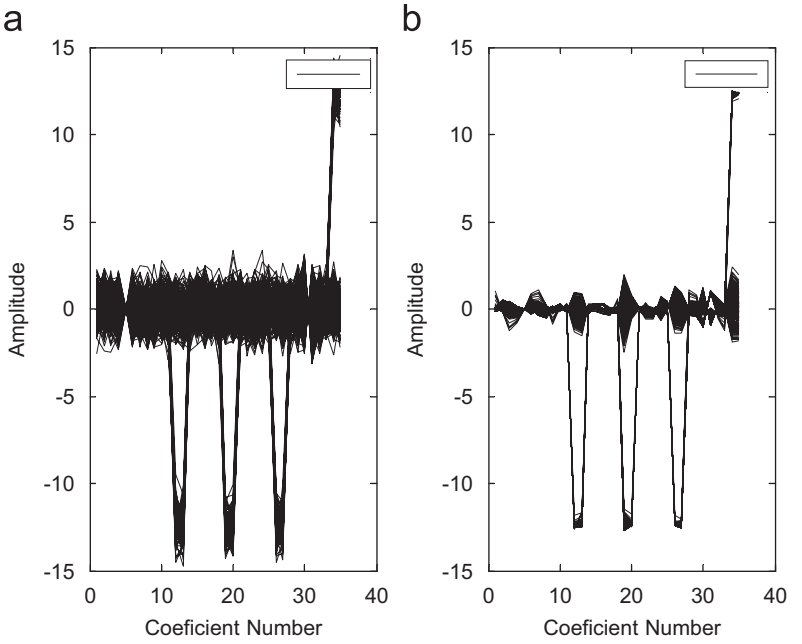

Fig. 5. Comparison of the AANN (a) input and (b) output wavelet coefficients corresponding to level 2 of decomposition.

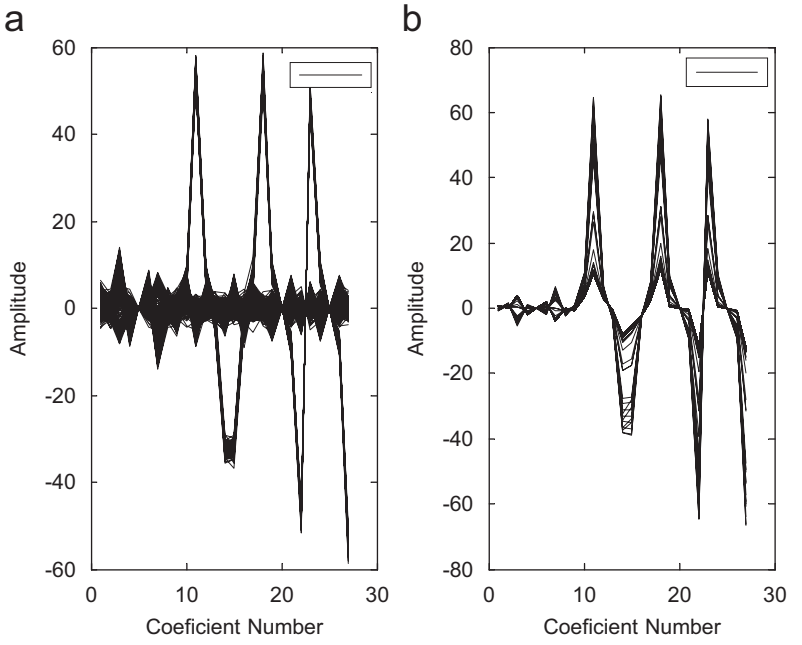

Fig. 6. Comparison of the AANN (a) input and (b) output wavelet coefficients corresponding to level 3 of decomposition.

In order to examine the sensitivity of the proposed method to detect faults, the 1404 signals generated by adding a perturbation to the healthy signals are tested. As remarked above, the perturbations consist of cubic splines added to the original signals at different locations between 0.1 and 0.4 Angular Units, with angular support lengths varying between 0.03 and 0.05 and with amplitudes included in the range 1–12.

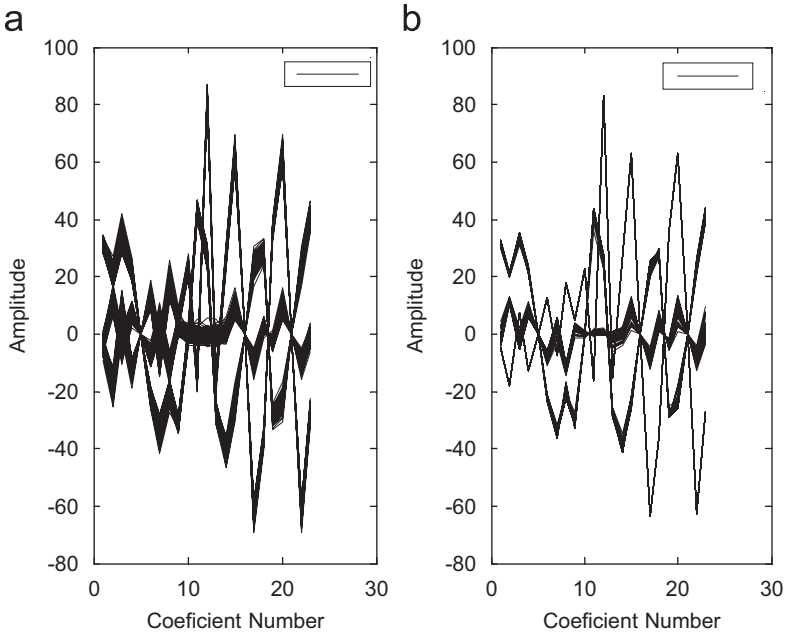

Fig. 7. Comparison of the AANN (a) input and (b) output wavelet coefficients corresponding to level 4 of decomposition.

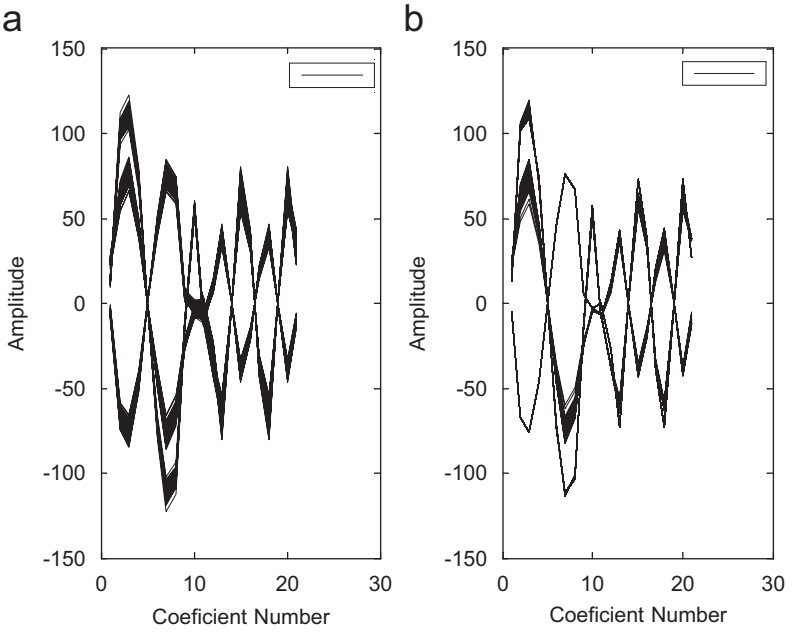

Fig. 8. Comparison of the AANN (a) input and (b) output wavelet coefficients corresponding to level 5 of decomposition.

The results obtained for the three signals, when different angular support lengths and amplitudes are considered, are shown in Tables 2 and 3, respectively. In these tables, the rate of successful damage detection is shown.

In the same way, to give an idea of the sensitivity of the novelty index to the amplitude of the perturbation, Fig. 9 shows the value of this index obtained for three training samples corresponding to perturbations of 9, 10 and 11 units of amplitude, respectively, centred at 0.25 units of angular shaft position and with supports of 0.03 angular units. In the same figure, the threshold values are represented too. According to Section 2.1, if

Table 1
Rate of false warnings of the trained AANNs

Level 2	6.50%
Level 3	2.50%
Level 4	4.50%
Level 5	2.50%

Table 2
Performance of the method for different angular support lengths of perturbation

Angular support	s1 (%)	s2 (%)	s3 (%)
0.03	70.51	75.54	80.13
0.04	83.33	82.69	84.61
0.05	83.33	88.46	85.25

Table 3
Performance of the method for different amplitudes of perturbation

Amplitude	s1 (%)	s2 (%)	s3 (%)
1	25.64	30.76	35.89
2	38.46	51.28	33.33
3	48.71	48.71	76.92
4	64.10	69.23	71.79
5	87.17	94.87	92.30
6	89.74	94.87	92.30
7	97.43	97.43	100
8	97.43	100	100
9	100	100	100
10	100	100	97.43
11	100	100	100
12	100	100	100

any component of the novelty index vector is higher than or equal to the threshold vector for any localisation and for any level of decomposition, the signal is considered to belong to a damaged machine. This is reasonable taking into account that a 4σ criterion was chosen in Eq. (2). Because of it, all the signals shown in Fig. 9 are considered to be damaged, observing the central region of the x-axis in which the perturbation was added to the original signal. In particular, for the perturbations of 9 and 10 units and level 5 of decomposition, points are on the threshold line near the central region, which means a fault signal. For the perturbation of 11 units and level 3 of decomposition, the point is clearly over the threshold line.

With the same purpose, the sensitivity of the novelty index to the length of the support was studied too. To do so, the novelty index obtained for three training samples, corresponding to a perturbation of 9 units of amplitude, centred at 0.25 units of angular shaft position, and with three different supports of 0.03, 0.04 and 0.05 angular units, respectively, for the perturbation, is plotted in Fig. 10. The same conclusions as in Fig. 9 can be obtained in this particular case.

Some observations may be made from Tables 2 and 3 and from Figs. 9 and 10 as follows:

(a) In general, the performance of the AANNs as a fault detection method is good.
(b) The length of the angular support used to define the perturbation, i.e., the duration of the perturbation, does not greatly affect the results.
(c) The sensitivity of the method depends on the amplitude of the perturbation, because when the level of the amplitude is close to the level of noise, the rate of success decreases. Nevertheless, the networks were capable of detecting about half of the damaged cases when the amplitude of the perturbation is equal to three.

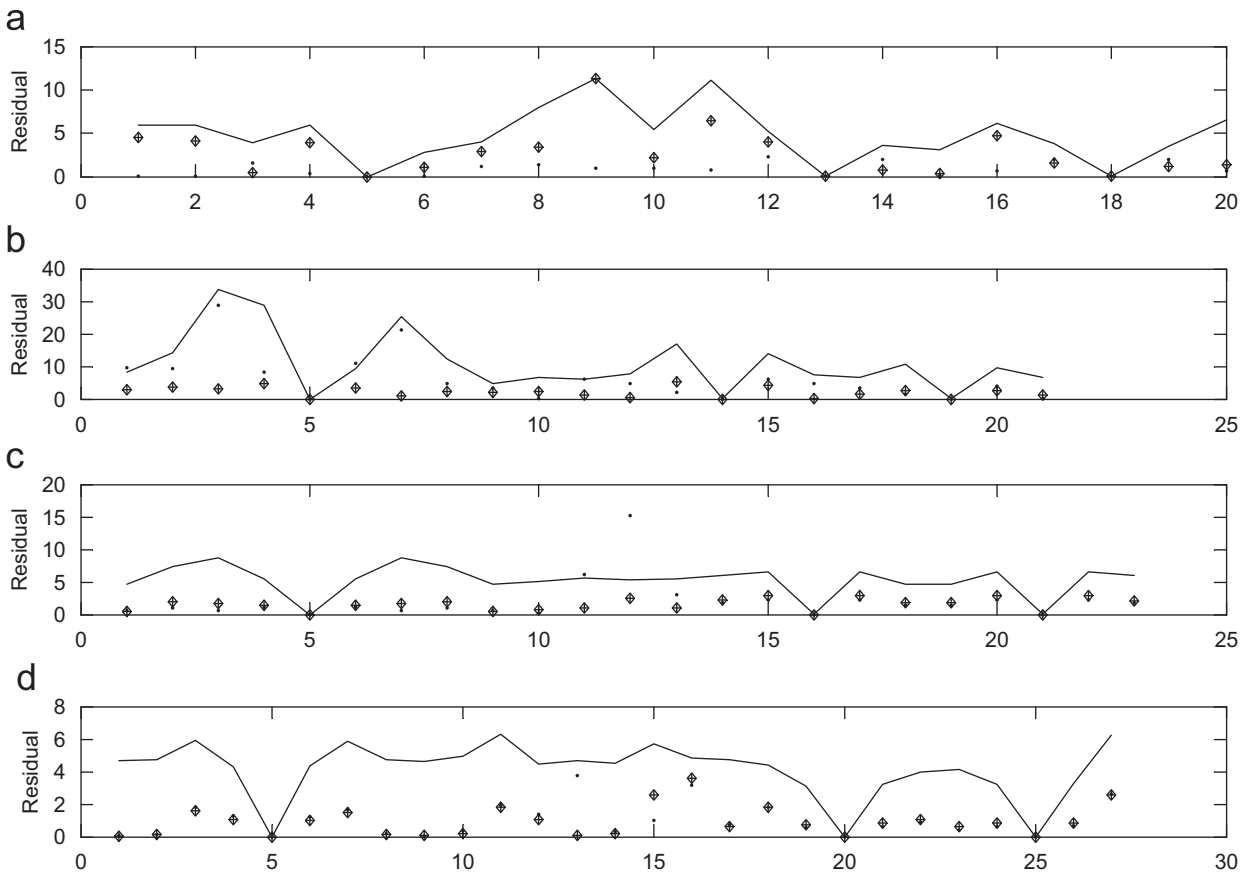

Fig. 9. Sensitivity of the novelty index vectors to the amplitude of perturbation: (a) Level 5 of decomposition, (b) level 4, (c) level 3, (d) level 2. With '◇' symbol are represented the samples where a perturbation of 9 units of amplitude is introduced; and the symbols '+' and '·' illustrate signals corrupted with perturbation of 10 and 11 units of amplitude, respectively. The continuous line represents the threshold values.

It is necessary to remark that, with this fault detection system a novelty index is defined for each wavelet coefficient obtained with the four levels of decomposition. This constitutes an advantage compared with other proposed detection methods since it yields a time–frequency distribution of the machine warnings that allows one to determine in which angular positions and frequencies the anomalies occur.

5.2. Fault diagnosis for pump rotor

5.2.1. Vibration data

The vibration signals for this second example were obtained from Ypma et al. [34]. They were measured from two identical pumps driven by an electric motor and composed of two delaying gear combinations. The number of teeth of the first set of gears was 20, while in the second set it was 40. While one of the pumps was operating in healthy conditions with no faults, the other was damaged due to the presence of pitting in both gears.

Acceleration vibration signals were measured with seven accelerometers located at different positions on the system. The first two were radially mounted near the driving shaft, separated at an angle of 90°, while the third accelerometer was used to measure the axial acceleration near the driving shaft; the remaining four accelerometers were installed radially at different locations on the machine casing.

The sets of measurements were obtained under two different operational conditions corresponding to high and low loads. Furthermore, the signals were low-pass filtered with cut-off frequencies of 1

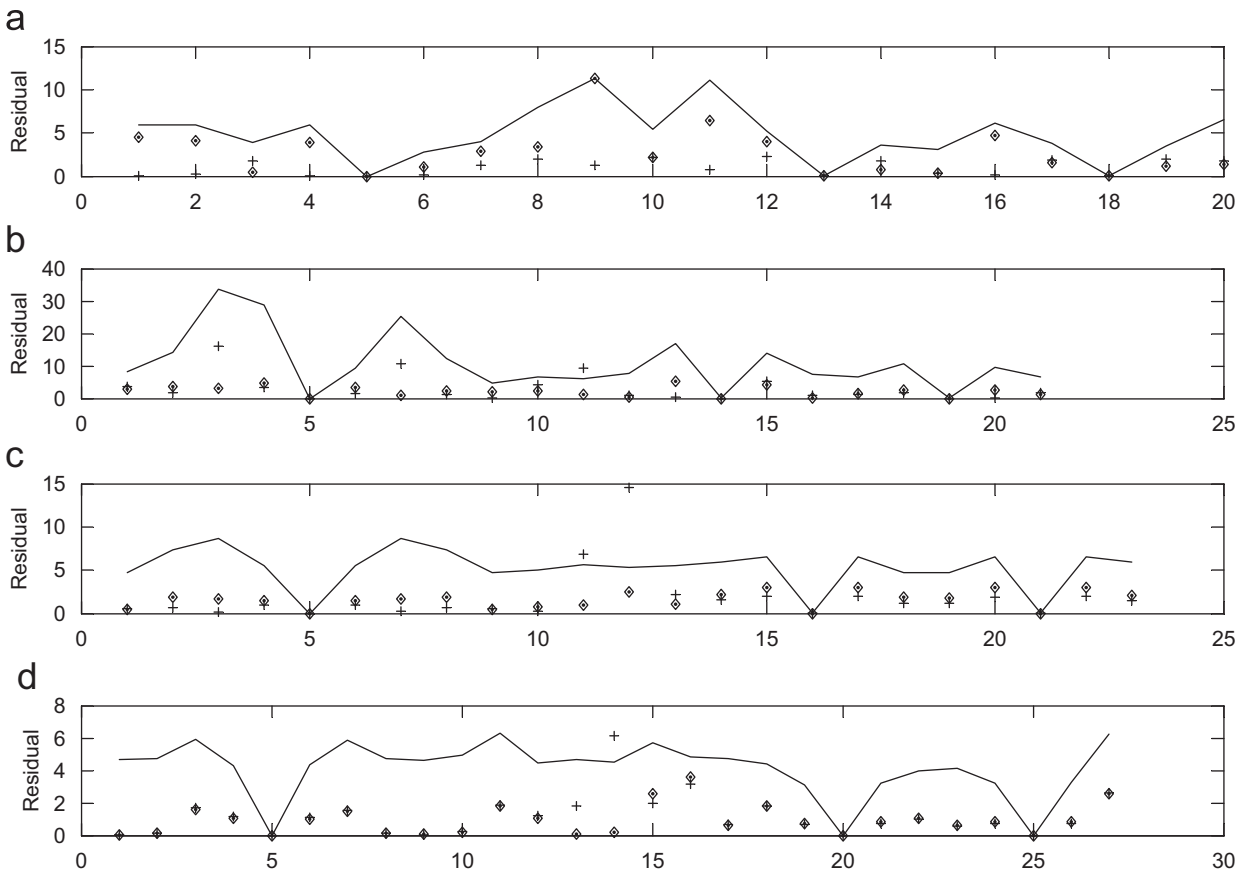

Fig. 10. Sensitivity of the novelty index vectors to the length of support of the perturbation: (a) Level 5 of decomposition, (b) level 4, (c) level 3, (d) level 2. With '◇' symbol are represented the samples where a perturbation of 0.05 units of angular support is introduced, while symbols '+' and '·' illustrate signals corrupted with perturbations of 0.03 and 0.04 units of angular support, respectively. The continuous line represents the threshold values.

and 5 kHz, respectively, for sampling rates of 3.2 and 12.8 kHz. Therefore, four sets of measurements were obtained.

In this work, only vibration signals measured with the filter of 1 kHz will be processed with DWT to extract the main features for using as inputs to AANNs according to the procedure proposed in Section 4. As the low-frequency measurements were registered during 24.3 s and the angular speed of the machine was 1000 rpm, 405 revolutions of the shaft were measured and, therefore, 405 signals were available for each sensor.

5.2.2. Signal pre-processing

As in the previous example, during this step it is necessary to select the most appropriate wavelet for the pre-processing as well as the level of decomposition. Gear damage will influence the vibrational measurements recorded in the machine casing by modulating the amplitudes of the signals at the meshing frequencies and their harmonics. This fact will help to establish the level of decomposition required.

To choose the optimal wavelet, only orthogonal and biorthogonal wavelets were considered as they enable the DWT to be carried out using the FWT. Haar wavelet was also eliminated because of its irregularity. Finally, although there were more candidates, Daubechies wavelet with six vanishing moments ('db6') was selected because of its capabilities.

On the other hand, as the meshing frequencies of both gear units are 333.3 and 666.6 Hz, respectively, three levels of decomposition of the selected wavelet are suitable for analysing the frequency affected by the damage. For each one of these levels, 105, 58 and 34 wavelet coefficients were obtained which were used as input of the NNs.

5.2.3. Training of the AANNs

In this example, it was necessary to establish 21 AANNs since vibration signals were measured on seven sensors and, besides, three levels of decomposition based on DWT were performed on every signal.

From the 810 signals recorded from the undamaged pump under the two levels of load, 650 were taken as training pattern for the AANNs, while the remaining 160 signals were used to test the training networks.

The training and configuration of the AANNs was performed as in the first example, the result being the topologies as shown in Table 4.

Once the AANNs were trained, the training patterns were introduced again in the networks, which allowed the residual vectors to be determined, and therefore, the novelty indexes.

Table 4
Topology of the different AANNs

	Level 1	Level 2	Level 3
Sensor 1	$105 \times 90 \times 1 \times 90 \times 105$	$58 \times 33 \times 1 \times 33 \times 58$	$34 \times 17 \times 1 \times 17 \times 34$
Sensor 2	$105 \times 45 \times 1 \times 45 \times 105$	$58 \times 31 \times 1 \times 31 \times 58$	$34 \times 18 \times 1 \times 18 \times 34$
Sensor 3	$105 \times 81 \times 1 \times 81 \times 105$	$58 \times 32 \times 1 \times 32 \times 58$	$34 \times 12 \times 1 \times 12 \times 34$
Sensor 4	$105 \times 59 \times 1 \times 59 \times 105$	$58 \times 37 \times 1 \times 37 \times 58$	$34 \times 16 \times 1 \times 16 \times 34$
Sensor 5	$105 \times 87 \times 1 \times 87 \times 105$	$58 \times 38 \times 1 \times 38 \times 58$	$34 \times 20 \times 1 \times 20 \times 34$
Sensor 6	$105 \times 86 \times 1 \times 86 \times 105$	$58 \times 37 \times 1 \times 37 \times 58$	$34 \times 20 \times 1 \times 20 \times 34$
Sensor 7	$105 \times 86 \times 1 \times 86 \times 105$	$58 \times 39 \times 1 \times 39 \times 58$	$34 \times 19 \times 1 \times 19 \times 34$

Table 5
Rate of false warnings of the trained AANNs

Low load		High load	
Level 1	Rate (%)	Level 1	Rate (%)
s1	3.95	s1	0.25%
s2	4.19	s2	13.82%
s3	3.95	s3	2.46%
s4	6.17	s4	8.88%
s5	2.46	s5	2.46%
s6	0	s6	4.19%
s7	6.17	s7	11.35%
Level 2	Rate (%)	Level 2	Rate (%)
s1	1.72	s1	11.60
s2	1.48	s2	1.48
s3	1.23	s3	0
s4	3.40	s4	0
s5	8.14	s5	6.91
s6	0.74	s6	10.37
s7	8.39	s7	10.12
Level 3	Rate (%)	Level 3	Rate (%)
s1	1.23	s1	0.74
s2	0.49	s2	0
s3	4.93	s3	0
s4	10.12	s4	6.17
s5	4.93	s5	4.93
s6	1.23	s6	0
s7	0.25	s7	0

5.2.4. Recognition testing

Firstly, the 160 signals, measured on the undamaged pump and not used during the training stage, were introduced in the AANNs to check the false damage warnings of the detection system. The rate of false warnings is shown in Table 5 for the two load conditions and for the seven sensors. The reliability of the networks is again very good when non-faulted signals are used.

Finally, the 810 signals recorded by each sensor over the damaged pump were introduced in the AANNs. The rate of successful damage detections is shown in Table 6.

From an observation of the results, it can be said that the diagnostic ability of the method is strongly sensitive to the transducer locations and to the operating conditions. Very good results were obtained for sensors 4, 6 and 7, radially mounted on the machine casing, for the first two levels of decomposition and under low load conditions. This variability can be due to the effect of the transfer function between the gear teeth and the casing location, where the vibration is picked up. According to this, the best locations would be those more sensitive to the frequencies of the fault and its harmonics. Moreover, this transfer function will also change with the operating conditions. Better results were obtained for the low level of load. This may be due to the fact that the working conditions under low load would be farther from the nominal conditions than those under high load which would contribute to increase in the amplitude of the vibrations. Actually, it would be advisable, for condition monitoring purposes, to compare vibration signals picked up under the same operating conditions, although it is practically impossible from a practical point of view.

On the other hand, although the method is suitable for localising faults, in this particular example, as the number of teeth on the input shaft is a multiple of the number of teeth on the output shaft, it is not possible to localise the component which fails and, therefore, only warning predictions are provided.

Table 6
Rate of success for damaged signals

Low load		High load	
Level 1	Rate (%)	Level 1	Rate (%)
s1	43.46	s1	0
s2	67.65	s2	7.65
s3	2.46	s3	3.70
s4	97.03	s4	63.20
s5	0.70	s5	0
s6	86.67	s6	0
s7	100	s7	100
Level 2	Rate (%)	Level 2	Rate (%)
s1	14.56	s1	0.25
s2	28.14	s2	3.20
s3	3.20	s3	0
s4	99.01	s4	2.71
s5	27.65	s5	13.82
s6	88.64	s6	5.18
s7	100	s7	83.20
Level 3	Rate (%)	Level 3	Rate (%)
s1	0.70	s1	0
s2	0.25	s2	1
s3	0	s3	0
s4	38.02	s4	5.67
s5	2.46	s5	0
s6	0.49	s6	0
s7	19.7%	s7	1.48

6. Conclusions

In this paper, a method for online fault diagnosis of rotating machinery, combining the capabilities of AANNs and WT, has been developed. Pattern recognition procedures based on neural approaches were used to compare DWT coefficients obtained from undamaged and damaged gear vibration signals. The difference between network output and input wavelet coefficients was used as a fault detection symptom. Furthermore, the use of a novelty index vector defined from DWT, unlike the scalar indexes used in other procedures, makes it possible to determine the ranges of time and frequency for which the faults appear, which can be used, in some cases, as a tool to localise the damaged tooth.

Another advantage of the proposed method is that DWT is performed directly on the raw vibration signals and, not by processing the TSA, whose evaluation can be complex. Then, as it is not necessary to calculate any ensemble average over many revolutions, relevant information is not lost by, for instance, lack of synchronism, and moreover, fault predictions can be given for every revolution, which makes the diagnosis procedure faster and requires less space to store the picked up signals.

On the other hand, according to the results presented, the analyses considered are very sensitive to the transducer locations as the machine casing acts as a filter whose transfer function is highly influenced by the spatial position of these transducers. Because of this, in order to optimise the locations, it would be very useful to perform a prior sensitivity study using an updated numerical model of the machine if it were available.

For future research, it is the intention of these authors to develop the next phase, which will allow this diagnosis procedure to be combined with another technique providing a criterion to quantify the severity of the fault.

Acknowledgements

The writers acknowledge support for the work reported in this paper from the Ministry of Education and Science of Spain (project BIA2004-06272).

The data set for example 2 was downloaded freely from the website: http://www.ph.tn.tudelft.nl/~ypma/mechanical.html. The authors gratefully acknowledge Dr. Alexander Ypma of Delft Technical University for making the data set available.

References

[1] P.D. McFadden, A revised model for the extraction of periodic waveforms by time domain averaging, *Mechanical Systems and Signal Processing* 1 (1987) 83–95.

[2] G. Dalpiaz, A. Rivola, R. Rubini, Effectiveness and sensitivity of vibration processing techniques for local fault detection in gears, *Mechanical Systems and Signal Processing* 14 (2000) 387–412.

[3] W.Q. Wang, F. Ismail, M.F. Golnaraghi, Assesment of gear damage monitoring techniques using vibration measurements, *Mechanical Systems and Signal Processing* 15 (2001) 905–922.

[4] W.J. Staszewski, K. Worden, G.R. Tomlinson, Time–frequency analysis in gearbox fault detection using the Wigner-Ville distribution and pattern recognition, *Mechanical Systems and Signal Processing* 11 (1997) 673–692.

[5] D. Boulahbal, M. Farid Golnaghari, F. Ismail, Amplitude and phase wavelet maps for the detection of cracks in geared systems, *Mechanical Systems and Signal Processing* 13 (1999) 423–436.

[6] D. Gabor, Theory of communication, *IEEE Journal* 93 (1946) 429–457.

[7] I. Daubechies, Orthonormal bases of compactly supported wavelets, *Communications on Pure and Applied Mathematics* XLI (1988) 909–996.

[8] S.G. Mallat, Theory for multiresolution signal decomposition: the wavelet representation, *IEEE Transactions on Pattern Analysis* 11 (1989) 674–693.

[9] I. Daubechies, The wavelet transfer, time–frequency localization and signal analysis, *IEEE Transactions on Information Theory* 36 (1990) 961–1005.

[10] W.J. Wang, P.D. McFadden, Application of wavelets to gearbox vibration signals for fault detection, *Journal of Sound and Vibration* 192 (1996) 927–939.

[11] Q. Wang, X. Deng, Damage detection with spatial wavelets, *International Journal of Solids and Structures* 36 (1999) 3443–3468.

[12] E. Douka, S. Loutridis, A. Trochidis, Crack identification in beams using wavelet analysis, *International Journal of Solids and Structures* 40 (2003) 3557–3569.

[13] A.V. Ovanesova, L.E. Suarez, Applications of wavelet transforms to damage detection in frame structures, *Engineering Structures* 26 (2004) 39–49.
[14] J. Han, W. Ren, Z. Sun, Wavelet packet based damage identification of beam structures, *International Journal of Solids and Structures* 42 (2005) 6610–6627.
[15] S.Q. Zhu, S.S. Lu, Wavelet-based crack identification of bridge beam from operational deflection time history, *International Journal of Solids and Structures* 43 (2006) 2299–2317.
[16] K. Gurley, A. Kareem, Application of wavelet transforms in earthquake, wind and ocean engineering, *Engineering Structures* 21 (1999) 149–167.
[17] G.G.W. Walter, *Wavelets and Other Orthogonal Systems with Applications*, CRC Press, Boca Raton, 1994.
[18] J.S. Walker, *A Primer on Wavelets and their Scientific Applications*, Chapman & Hall/CRC, London/Boca Raton, FL, 1999.
[19] S.G. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, New York, 1999.
[20] B.A. Paya, I.I. Esat, M.N.M. Badi, Artificial neural network based fault diagnostics of rotating machinery using wavelet transforms as a preprocessor, *Mechanical Systems and Signal Processing* 11 (1997) 751–765.
[21] R.R.K. Reddy, R. Ganguli, Structural damage detection in a helicopter rotor blade using radial basis function neural networks, *Smart Materials Structures* 12 (2003) 232–241.
[22] B.S. Yang, T. Han, J.L. An, ART-KOHONEN neural network for fault diagnosis of rotating machinery, *Mechanical Systems and Signal Processing* 18 (2004) 645–657.
[23] B. Samanta, Gear fault detection using artificial neural networks and support vector machines with genetic algorithms, *Mechanical Systems and Signal Processing* 18 (2004) 625–644.
[24] C.M. Bishop, Novelty detection and neural network validation, *IEE Proceedings on Vision and Image Signal Processing* 141 (1994) 217–222.
[25] K. Worden, Structural fault detection using a novelty measure, *Journal of Sound and Vibration* 201 (1997) 85–101.
[26] D.J. Chen, W.J. Wang, Pattern classification of wavelet map using multi-layer perception neural network for gear fault detection, *Mechanical Systems and Signal Processing* 16 (2002) 695–704.
[27] A. Rytter, Vibration based Inspection of Civil Engineering Structures, PhD Thesis, Department of Building Technology and Structural Engineering, University of Aalborg, Denmark, 1993.
[28] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.
[29] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, *IEEE Transactions on Pattern Analysis* 22 (2000) 4–37.
[30] S. Geman, E. Bienestock, R. Doursat, Neural networks and the bias/variance dilemma, *Neural Computation* 4 (1992) 1–58.
[31] G. Strang, T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Wellesley, MA, 1996.
[32] I. Daubechies, *Ten Lectures on Wavelets*, Rutger University and AT&T Laboratories, 1995.
[33] M. Verteli, J. Kovacevic, *Wavelets and Subband Coding*, Pearson Education, 1995.
[34] A. Ypma, R.P.W. Ligteringen, E.E.E. Duin, E.E.E. Frietman, *Pump vibration datasets*, Pattern recognition group, Delf University of Techology, 1999.