Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Mechanical Systems and Signal Processing 18 (2004) 1077–1095
Mechanical Systems and Signal Processing
www.elsevier.com/locate/jnlabr/ymssp

Bearing fault diagnosis based on wavelet transform and fuzzy inference

Xinsheng Lou[1], Kenneth A. Loparo*

Department of Electrical and Computer Science, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA

Received 12 May 2003; received in revised form 19 May 2003; accepted 22 May 2003

Abstract

This paper deals with a new scheme for the diagnosis of localised defects in ball bearings based on the wavelet transform and neuro-fuzzy classification. Vibration signals for normal bearings, bearings with inner race faults and ball faults were acquired from a motor-driven experimental system. The wavelet transform was used to process the accelerometer signals and to generate feature vectors. An adaptive neural-fuzzy inference system (ANFIS) was trained and used as a diagnostic classifier. For comparison purposes, the Euclidean vector distance method as well as the vector correlation coefficient method were also investigated. The results demonstrate that the developed diagnostic method can reliably separate different fault conditions under the presence of load variations.

Keywords: Wavelets; Fault diagnosis; Fuzzy inference; Pattern classification; Bearings

1. Introduction

Condition monitoring of rotating machinery is important in terms of system maintenance and process automation. Rolling element bearing failures are one of the foremost causes of failures in rotating machinery. This necessitates the development, implementation, and deployment of on-line diagnostic monitoring systems that are independent of operating conditions.

In most machine fault diagnosis and prognosis systems, the vibration of the rotating machine (motor, gearbox, etc.) is directly measured by an accelerometer, in some few cases, by an acoustic pickup. Some techniques use the stator currents of the electrical motor as the input signals for

*Corresponding author. Tel.: +1-216-368-4115; fax: +1-216-368-3123.
E-mail address: kal4@po.cwru.edu (K.A. Loparo).
[1] Current affiliation: E&ES Department, the Alstom Power Plant Laboratories, Windsor, CT 06095, USA.

0888-3270/03/$ - see front matter
doi:10.1016/S0888-3270(03)00077-3

fault detection [1]. Fault signal detection and recognition are often accomplished by pattern recognition using a neural network [2,3], RBF network [4], Gaussian mixture model network [5,6], fuzzy logic network [5], Bayesian classifier [7], vector correlation or vector distance measure [8]. Commonly used feature generation methods include the short-time Fourier transform (STFT) [2], wavelet time-scale decomposition [2,9,10], cumulant spectrum [8], etc.

The discrete wavelet transform (DWT) provides an efficient method for generating feature vectors. The DWT coefficients can be used to generate statistical parameters from each resolution level of the transform. This method of feature extraction has been used to recognise signals from RF transmitters with a back propagation neural network [9] and ground vehicles with vector correlation and distance pattern matching [11]. Acoustic analysis methods have been developed to detect and classify underwater objects using wavelets with a neural network and quadratic Bayesian classifiers [12]. The discriminative feature extraction recogniser, which combines a feature extractor and classifier, is presented in [13]. This network optimises both a feature extraction process and a classification process by pattern production and adaptation. As an alternative to the back propagation neural network, a supervised radial basis function network is used. A new network type called "wave-net" [14] adapts the RBF network concept, and uses wavelets as the basis functions for the network. This network has been used for speaker identification [15]. Liu and Ling have applied the principle of mutual information to the identification of wavelets that carry significant information of machinery faults, instead of the "best matching" criterion used in matching pursuit [16]. Altmann and Mathew have used ANFIS for automated selection of wavelet packets containing bearing fault related features [17]. Peng et al., proposed a fusion fault diagnosis method based on the wavelet transform, genetic algorithms and neural networks [10]. Xu and Chan have done very similar work [18].

In this paper, a new technique for localised bearing fault diagnosis is developed using the discrete wavelet transform (DWT). In this method, experimental vibration signals for normal and faulty bearings are pre-processed to obtain a (0,1) normal distribution where the wavelet transform was used to process the normalised data. Then a feature vector is defined using the components from the DWT. By using selected segments from the available experimental data, typical sample feature vectors are generated for both normal bearings and bearings with different types of faults under different load conditions. Then different pattern classification methods have been studied in the decision making stage, including the neural-fuzzy inference system, which is believed to be most suitable for complex situations due to its adaptability and the capability of the network to realise a non-linear approximation.

2. Experimental system

The ball bearings are installed in a motor driven mechanical system, as shown in Fig. 1. A 2 hp, three-phase induction motor (Reliance Electric 2HP IQPreAlert motor), was connected to a dynamometer and a torque sensor by a self-aligning coupling. The dynamometer is controlled so that desired torque load levels can be achieved. An accelerometer with a bandwidth up to 5000 Hz and a 1 V/g output is mounted on the motor housing at the drive-end of the motor to acquire the vibration signals from the bearing. The data collection system consists of a high bandwidth amplifier particularly designed for vibration signals and a data recorder with a sampling frequency

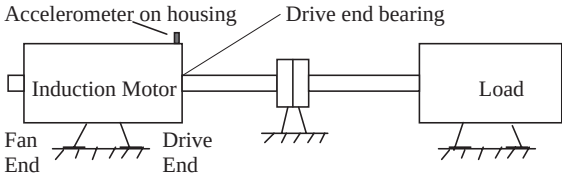

Fig. 1. A schematic of the experimental system.

of 12,000 Hz per channel. The data recorder is equipped with low-pass filters at the input stage for anti-aliasing. On the other hand, the frequency content of interest in the vibration signals of the system under study did not exceed 5000 Hz, for which the sampling rate is ample.

To develop the new diagnostic technique, four sets of data were obtained from the experimental system (shown in Fig. 1): (i) under normal conditions; (ii) with inner race faults; (iii) with a ball fault (iv) with outer race faults. Faults were introduced into the drive-end bearing of the motor using the EDM method.

The bearings used in this work are deep grove ball bearing manufactured by NTN. Some parameters are listed below:

Bearing specs, NTN p/n 6205c3:
Basic dynamic load rating: 14000 N
Basic static load rating: 7850 N
Radial internal clearance: 0.013–0.028
Pitch diameter(Pd): 1.535 in
Ball diameter(Bd): 0.312 in
Ball pass frequency at outer ring(OR): $3.59 \times$ rps
Ball pass frequency at outer ring(IN): $5.41 \times$ rps
Fundamental train frequency (FTF): $0.40 \times$ rps
Ball spin (BS) frequency: $2.36 \times$ rps
In the experiments, rps ≈ 30 Hz, for zero load, which yields:
FTF $\approx 0.40 \times 30 = 12$ Hz; BS $\approx 2.36 \times 30 = 70.8$ Hz;
OR $\approx 3.59 \times 30 = 107.7$ Hz; IR $\approx 5.41 \times 30 = 162.3$ Hz.

The sizes of the defects for the NTN bearing described previously are:
Inner race defect size: diameter = 40 mils, depth = 40 mils;
Outer race defect size: diameter = 40 mils, depth = 40 mils;
Ball defect size: diameter = 40 mils, depth = 40 mils.

Each bearing is tested under four different loads (0, 1, 2 and 3 hp). Frequency domain analysis was performed using the DTFT and then a fault detection filter was designed to separate the normal and faulty modes [19]. The data collected for the outer race defected bearings was found to be corrupted and this data was not used in the subsequent analysis.

In this paper wavelet analysis was used to process the test data, and methods were developed to separate the three classes of data: normal, ball fault and inner race fault. These techniques enable the detection of abnormalities in the bearing and at the same time identification of the type of a fault.

3. Preprocessing test data

By examining the magnitude of the vibration data under operating conditions with severe bearing faults, it is possible to distinguish the normal data from different types of fault data. However, this is not always applicable because the signal morphology that results from a fault changes over time as the fault progresses from initiation to failure. Thus, some faults will be undetectable until failure is imminent. Because the early detection and isolation of faults is important for condition-based maintenance, a more sophisticated signal processing approach is necessary. To accomplish this objective, we need to carefully examine the signals. The first step in our approach is to preprocess the test data before performing the wavelet analysis.

To make the signals comparable regardless of differences in magnitude, the signals are normalised by using the following equation:

$$s_{p_i} = \frac{s_i - \mu}{\sigma}, \tag{1}$$

where s_i is the ith element of the signal (column vector) S, μ and σ are the mean and standard deviation of the vector S, respectively; s_{pi} is the ith element of the signal series S_p after normalisation.

The assumption of Eq. (1) is that the signals have a normal probability distribution (or are at least close to Gaussian). From the histograms in Fig. 2, it can be seen that the signals all have a single peak and appear to be approximately normally distributed. To confirm this observation, the χ^2 test of goodness-of-fit (Table 1) is used [20]. The results show that data for normal operating conditions and inner race fault conditions fit a normal distribution very well, while the signals for ball fault conditions fit this hypothesis to a lesser degree. However, it will be seen in Section 5 that this will not have a significant influence on the proposed diagnosis method as long as its distribution is close enough to a normal probability distribution so that normalisation using Eq. (1) introduces insignificant changes to the statistical signatures of the signals.

Fig. 3 shows a comparison of the preprocessed data for the three different types of vibration data: normal, inner race fault and ball fault.

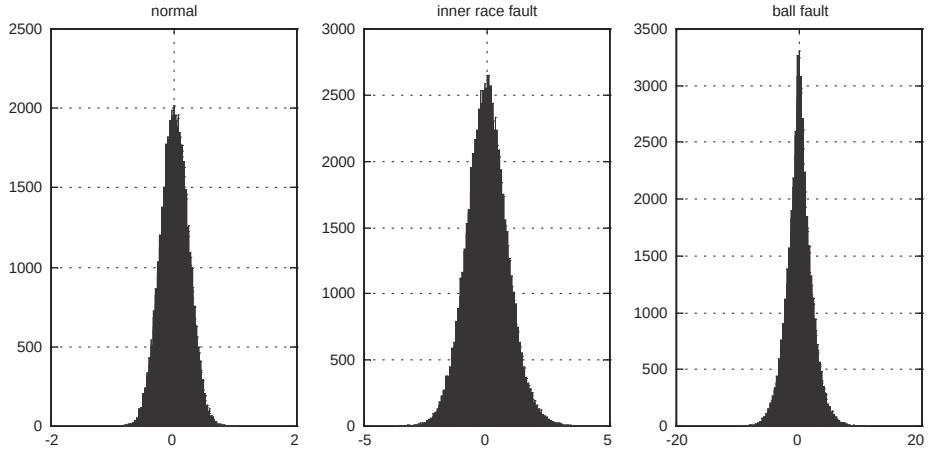

Fig. 2. Histograms of the test data.

Table 1
Results of χ^2 test

Results	0 hp		1 hp		2 hp		3 hp	
	χ^2	*P*-Value	χ^2	*P*-Value	χ^2	*P*-Value	χ^2	*P*-Value
Normal	19.8753	0.06848	16.4985	0.1695	5.8771	0.9222	10.3054	0.5892
Inner race fault	8.2899	0.7621	7.1593	0.8469	5.2120	0.9505	13.3171	0.3464
Ball fault	20.5056	0.05811	45.5348	8.34e-6	23.392	0.02459	30.7598	0.002143

Note: 13 bins are used in the histogram, i.e. the freedom in χ^2 distribution is 12. Those with *P*-values larger than 0.05 are considered to be represented to some degree by a Normal probability distribution.

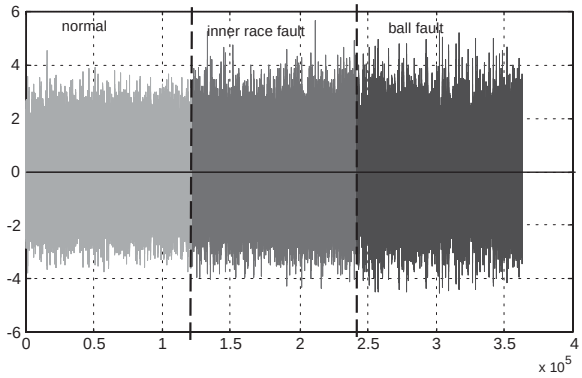

Fig. 3. Comparison of the preprocessed data of three different types of vibration data.

4. Wavelet analysis and feature extraction

4.1. Brief review of the wavelet theory

The Wavelet Transform is defined as the integral of the signal $s(t)$ multiplied by scaled, shifted versions of a *basic wavelet* function $\psi(t)$—a real-valued function whose Fourier Transform satisfies the admissibility criteria [21–23]:

$$C(a,b) = \int_R s(t)\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)\mathrm{d}t, \quad a \in R^+ - \{0\}, \quad b \in R. \tag{2}$$

where a is the so-called scaling parameter, b is the time localisation parameter. Both a and b can be continuous or discrete variables.

Multiplying each coefficient by an appropriately scaled and shifted wavelet yields the constituent wavelets of the original signal. For signals of finite energy, continuous wavelet synthesis provides the reconstruction formula:

$$s(t) = \frac{1}{K_\psi}\int_R\int_{R^+} C(a,b)\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)\frac{\mathrm{d}a}{a^2}\mathrm{d}b. \tag{3}$$

Associated with the wavelet ψ, which is used to define the details (high scale/low frequency content) in the decomposition, a scaling function ϕ, is used to define the approximations (low scale/high frequency content). Note $\int \phi(x)\,dx = 1$ while $\int \psi(x)\,dx = 0$.

To avoid intractable computations when operating at every scale of the CWT, scales and positions can be chosen based on a power of two, i.e. dyadic scales and positions. The discrete wavelet transform (DWT) analysis is more efficient and just as accurate. In this scheme, a and b are given by:

$$(j,k)\in Z^2: \quad a = 2^j, \quad b = k2^j, \quad Z = \{0, \pm 1, \pm 2, \ldots\}.$$

Let us define:

$$(j,k)\in Z^2: \quad \psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k), \quad \phi_{j,k}(t) = 2^{-j/2}\phi(2^{-j}t - k).$$

A wavelet filter with impulse g, plays the role of the wavelet ψ, and a scaling filter with impulse response h, plays the role of scaling function ϕ. g and h are defined on a regular grid ΔZ, where Δ is the sampling period (here, without loss of generality, set $\Delta = 1$). Then the discrete wavelet analysis can be described mathematically as:

$$C(a,b) = c(j,k) = \sum_{n\in Z} s(n)g_{j,k}(n),$$
$$a = 2^j, \quad b = k2^j, \quad j\in N, k\in N. \tag{4}$$

And discrete synthesis:

$$s(t) = \sum_{j\in Z}\sum_{k\in Z} c(j,k)\psi_{j,k}(t). \tag{5}$$

The detail at level j is defined as:

$$D_j(t) = \sum_{k\in Z} c(j,k)\psi_{j,k}(t) \tag{6}$$

and the approximation at level J:

$$A_{J-1} = \sum_{j>J} D_j. \tag{7}$$

Obviously, the following equations hold:

$$A_{j-1} = A_j + D_j, \tag{8}$$

$$s = A_j + \sum_{j\leqslant J} D_j. \tag{9}$$

In practice, the decomposition can be determined iteratively, with successive approximations being computed in turn, so that a signal is decomposed into many lower-resolution components. This is known as the *wavelet decomposition tree*. By using reconstruction filters and upsampling, we can reconstruct the signal constituents at each level of the decomposition [21–23].

Ingrid Daubechies invented what are called 'compactly supported orthonormal wavelets'—thus making discrete wavelet analysis practical. These wavelets have no explicit expression except for 'Daubechies-1 wavelet', which is the *Haar* wavelet. However, the square modulus of the transfer

function of h is explicit and fairly simple [22]. In this research work, 'Daubechies-2' and 'Daubechies-10' wavelets were used for signal processing and analysis.

4.2. Wavelet analysis and feature definition

We begin by examining the first half of the signals ($2^{16} = 65{,}536$ points) for analysis, and leave the rest of the signals for testing the method that we are developing. After more experimentation, the Daubichies-2 wavelet was selected for signal analysis and synthesis.

Fig. 4 shows a combination of signals for normal operating conditions but with different load conditions (0 and 3 hp); Fig. 5 shows the comparison of a signal under the normal operating condition and a signal under an inner race fault (both with 0 hp load). In these figures, the approximation (a_5), and five levels of details ($d_1 - d_5$) are chosen for each signal. Fig. 6 shows a comparison between a normal operating condition and a ball fault condition. From these plots, it appears that the inner race fault can be separated from the normal condition because, for example, d_3 and d_4 are quite different in magnitude between the two conditions; and for a ball fault $d_3 - d_4$ and a_5 are much smaller than those of the test data under the normal operating condition.

Using histograms, it is observed that the details and the approximation of a test signal still have a probability distribution that is close to a Normal distribution with zero mean. To quantify the features extracted using the wavelet decomposition, we define a DWT feature vector for a given

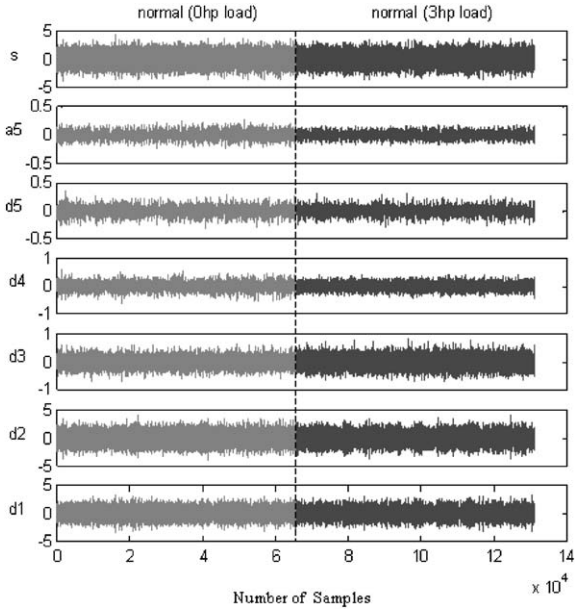

Fig. 4. Decomposition of signals under normal operating conditions (s—the original signal; a_5—the 5th level approximation, frequency range: 0–188 Hz; $d_1 - d_5$: the five details, d_1 (3000–6000 Hz), d_2 (1500–3000 Hz), d_3 (750–1500 Hz), d_4 (375–750 Hz), d_5 (188–375 Hz)).

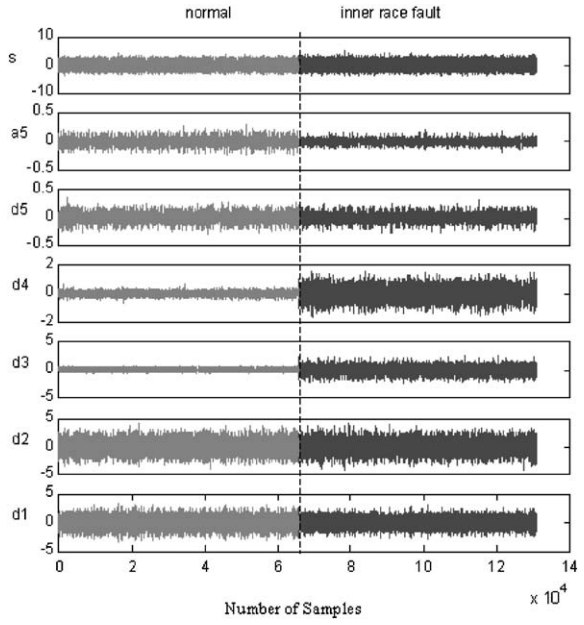

Fig. 5. Decomposition of signals—normal/with inner race fault (s—the original signal; a_5—the 5th level approximation, frequency range: 0–188 Hz; $d_1 - d_5$: the five details, d_1 (3000–6000 Hz), d_2 (1500–3000 Hz), d_3 (750–1500 Hz), d_4 (375–750 Hz), d_5 (188–375 Hz)).

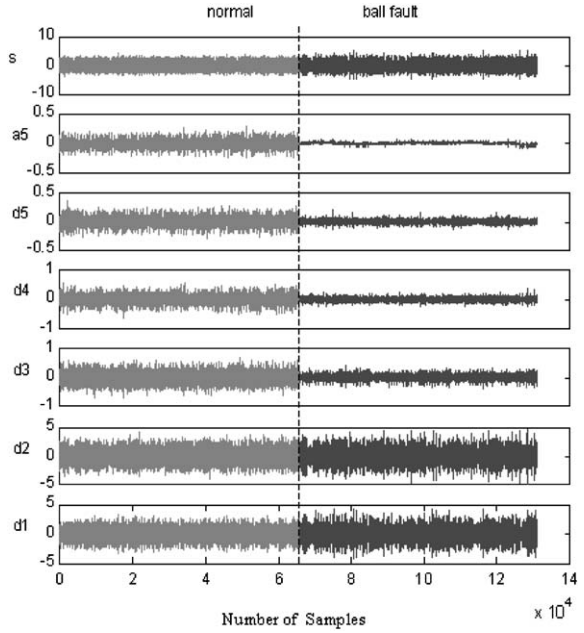

Fig. 6. Decomposition of signals—normal/with ball fault (s—the original signal; a_5—the 5th level approximation, frequency range: 0–188 Hz; $d_1 - d_5$: the five details, d_1 (3000–6000 Hz), d_2 (1500–3000 Hz), d_3 (750–1500 Hz), d_4 (375–750 Hz), d_5 (188–375 Hz)).

signal as $v = [v_1, v_2, \ldots, v_6]^T$ with its element defined as:

$$v_i = \sigma_i / \sigma_{ri}, \tag{10}$$

where $i = 1, \ldots, 6$, corresponds to $d_1, d_2, \ldots, d_5, a_5$, respectively and σ_i is the standard deviation of the ith decomposition , e.g. σ_1 is the standard deviation of d_1; σ_{ri} is the standard deviation of ith decomposition of a reference signal (in this case we have chosen a data set acquired under normal operating condition and 0 hp load). Note that the standard deviation used here is equivalent to the root mean square average of the signal because the signal has zero mean. This is used to quantify the average energy level of a signal.

4.3. Feature vector formation

The DWT feature vectors, as defined, are calculated for the test data, and listed as ($v_1 - v_6$) in Tables 2–4. The vd_k and cr_k are the vector distance and the correlation coefficient, as defined in the next section.

From the above analysis, a fast fault detection scheme can be developed by using the standard deviations of the wavelet decompositions d_3 and d_4 in a moving window as fault indicators. Thresholds can be chosen easily (for example, the upper threshold can be chosen in the range 1.2–3.2, and the lower threshold can be chosen in the range 0.4–0.8). Of course, we can also make further inferences on the fault type based on whether they are above the upper threshold or below the lower threshold. However, more objective methods are needed for making diagnostic conclusions. In the next section, vector distance and correlation coefficient methods are tested and investigations will be carried out using a neural fuzzy inference technique to ensure reliable diagnostic decisions in case the clusters are not geometrically distinct or linear correlation is no longer a valid method.

Table 2
Vectors and clustering results—normal conditions

v	0 hp	1 hp	2 hp	3 hp	Average
v_1	0.9968	1.0014	1.0039	0.9984	1.0001
v_2	1.0027	1.0032	0.9998	0.9990	1.0012
v_3	1.0005	0.8899	0.9906	1.1651	1.0115
v_4	0.9857	0.9142	0.8891	0.9163	0.9263
v_5	1.0024	1.0209	0.9609	0.9184	0.9757
v_6	0.9934	0.9226	0.8736	0.8556	0.9113
vd_1	**0.0111**	**0.0171**	**0.0035**	**0.0301**	**0**
vd_2	17.9471	19.0012	18.6185	17.5451	18.2629
vd_3	1.6564	1.3842	1.3278	1.5212	1.457
cr_1	**0.7595**	0.3566	**0.9855**	**0.8133**	**1**
cr_2	−0.5347	**−0.7507**	−0.2543	0.3586	−0.1277
cr_3	0.3368	0.5724	0.7121	0.2549	0.6066

Note: For 0 hp load, we use a different segment other than the segment used as a reference, therefore v_i are not exactly 1.

Table 3
Vectors and clustering results—inner race fault

v	0 hp	1 hp	2 hp	3 hp	Average
v_1	0.8061	0.8014	0.7262	0.6688	0.7506
v_2	0.9253	0.8771	0.8396	0.7803	0.8556
v_3	3.2805	3.4633	3.7122	3.9320	3.5970
v_4	3.3564	3.8875	4.6590	5.3547	4.3144
v_5	0.9119	0.8702	0.8578	0.9481	0.8970
v_6	0.6393	0.6979	0.8558	1.0673	0.8151
vd_1	11.1752	14.8915	21.3448	28.3238	18.2629
vd_2	**1.0570**	**0.2176**	**0.1360**	**1.2730**	**0**
vd_3	17.7693	22.3319	30.1552	38.5713	26.536
cr_1	0.0007	−0.0681	−0.1569	−0.2271	−0.1277
cr_2	**0.9903**	**0.9981**	**0.9994**	**0.9944**	**1**
cr_3	−0.4083	−0.4242	−0.4575	−0.4948	−0.4536

Table 4
Vectors and clustering results—ball fault

v	0 hp	1 hp	2 hp	3 hp	Average
v_1	1.0507	1.0760	1.0760	1.0863	1.0722
v_2	0.9826	0.9618	0.9623	0.9535	0.9650
v_3	0.4208	0.4184	0.3966	0.4041	0.4100
v_4	0.3869	0.3730	0.3291	0.3169	0.3515
v_5	0.3765	0.3542	0.3308	0.3351	0.3492
v_6	0.3162	0.3087	0.2988	0.3028	0.3066
vd_1	1.3559	1.4146	1.5331	1.5307	1.4569
vd_2	26.1389	26.3062	26.8295	26.8752	26.5356
vd_3	**0.003**	**0.0006**	**0.0011**	**0.0018**	**0**
cr_1	0.6056	0.6042	0.6039	0.6113	0.6066
cr_2	−0.4432	−0.4378	−0.4613	−0.4698	−0.4535
cr_3	**0.999**	**0.9998**	**0.9999**	**0.9993**	**1**

5. The decision making processes

5.1. Vector distance and correlation coefficient

As seen in the previous tables, two simple pattern classification methods can be used to make diagnostic decisions: (1) the Euclidean vector distance, and (2) the vector correlation coefficient method.

5.1.1. Euclidean vector distance

The average vector can be taken as the geometric centre of a particular cluster. Thus, when a feature vector has been obtained using the above method, the distances from the vector (or a point in a six-dimensional space) to the three centres can be calculated. For simplicity, we choose the square of the Euclidean distance as the distance metric:

$$vd_k = d_k^2 = \sum_{i=1}^{6} (v_{\text{test}} - v_{ck})^2, \quad k = 1, 2, 3, \tag{11}$$

where v_{cki} is the ith element of centre vector v_{ck}, $v_{\text{test}\,i}$ is the ith element of vector v_{test}, which is to be classified; $k = 1, 2, 3$ corresponds to the three geometric centres, respectively. The rule is that the smallest vector distance corresponds to the cluster that the given vector should belong to. Besides Euclidean Distance, Mahalanobis Distance can also be used for class separation, which scales the Euclidean distance by the covariance matrix [21].

5.1.2. Vector correlation coefficient

For computing the correlation coefficients, we can consider two feature vectors as a pair of random variables x and y. The correlation coefficient of v_{test} and v_{cr} is defined as:

$$cr_k = \frac{\text{cov}[v_{\text{test}} v_{ck}]}{\sigma_{v_{\text{test}}} \sigma_{v_{ck}}}, \quad k = 1, 2, 3, \tag{12}$$

where cov denotes the covariance of the two vectors, σ indicates the standard deviation. The rule is that the largest correlation coefficient for the unknown feature vector provides the fault type.

The results obtained by applying the above two methods to the feature vectors have also been listed in Tables 2–4. The results that were obtained by using the unused portion of the test data are listed in Tables 5–7.

The results using the first method are listed in the tables as (vd_1, vd_2, vd_3). Fortunately, in this case, the three clusters do not overlap, as can be seen from the values of the vector distances given

Table 5
Testing results—normal

v	0 hp	1 hp	2 hp	3 hp	Mixed
v_1	0.9995	1.0024	1.0058	1.002	1.0025
v_2	1.0004	1.0021	0.9974	0.9949	0.9982
v_3	1.0134	0.8951	1.0125	1.1807	1.0409
v_4	0.9843	0.9137	0.9075	0.9711	0.9603
v_5	0.9917	1.0274	0.9703	0.9086	0.9973
v_6	0.9758	0.9522	0.8955	0.8798	0.8913
vd_1	**0.0078**	**0.0181**	**0.0007**	**0.0362**	**0.0029**
vd_2	17.8819	18.9858	18.3838	17.103	17.8831
vd_3	1.6318	1.4344	1.4102	1.6253	1.5366
cr_1	**0.9428**	**0.2405**	**0.9974**	**0.6825**	**0.9157**
cr_2	0.1731	**−0.8536**	−0.1649	0.5349	0.2161
cr_3	0.4171	0.4896	0.624	0.1565	0.3672

Table 6
Testing results—inner race fault

v	0 hp	1 hp	2 hp	3 hp	Mixed
v_1	0.8045	0.7929	0.7345	0.6623	0.751
v_2	0.9256	0.8855	0.8288	0.7784	0.8592
v_3	3.2951	3.4507	3.7345	3.9307	3.5488
v_4	3.3469	3.8702	4.6884	5.4126	4.4364
v_5	0.9379	0.8512	0.9435	0.9392	0.9204
v_6	0.6147	0.6962	0.7675	1.069	0.9171
vd_1	11.2076	14.7344	21.6905	28.8381	18.8439
vd_2	**1.0768**	**0.2376**	**0.1642**	**1.3973**	**0.0282**
vd_3	17.8111	22.1154	30.5596	39.1417	27.3519
cr_1	0.01	−0.0674	−0.1419	−0.2318	−0.1649
cr_2	**0.9889**	**0.998**	**0.9994**	**0.9938**	**0.9992**
cr_3	−0.4076	−0.4218	−0.4556	−0.4941	−0.4683

Table 7
Testing results—ball fault

v	0 hp	1 hp	2 hp	3 hp	Mixed
v_1	1.0567	1.058	1.0968	1.0834	1.0832
v_2	0.9777	0.9767	0.9442	0.956	0.9553
v_3	0.4254	0.4042	0.3943	0.4005	0.445
v_4	0.3848	0.3794	0.338	0.3181	0.3714
v_5	0.3904	0.3719	0.3306	0.321	0.3475
v_6	0.2478	0.384	0.3324	0.3164	0.3493
vd_1	1.4233	1.3145	1.4909	1.5348	1.3483
vd_2	26.1874	26.2491	26.751	26.8895	26.1217
vd_3	**0.0069**	**0.0077**	**0.0025**	**0.0023**	**0.0037**
cr_1	0.6475	0.5536	0.5773	0.5972	0.5981
cr_2	−0.4079	−0.4947	−0.4739	−0.4712	−0.4453
cr_3	0.9943	0.9969	0.9981	0.9989	0.9987

in Tables 2–4. The testing results provide further evidence that the three clusters are separated, as shown in Tables 5–7. The test data uses $2^{15} = 32{,}768$ points from the unused portion of the experimental data. The 2^{15} data points are a combination of four parts (each has a length of 2^{13}), where each part contains data from one of the four possible load conditions.

The results using the second method are listed in the tables as (cr_1, cr_2, cr_3). By looking for the maximum of the correlation coefficients, two of the fault types can be reliably separated. However, some feature vectors for the normal operating condition are mis-classified. Nevertheless, this is an effective method for fault diagnosis when the fault has been detected using some other methods such as the fault detection filter that was developed in [19].

Other tests have also been carried out. For example, using only 4096 data points in a test vector and applying both of the above decision making methods. It is notable that the results are the same as the tests using 32,768 points: the vector distance method works very well in classifying the three clusters; and the correlation coefficient method can reliably separate inner race faults from ball faults. Note that this will significantly reduce the computation time and hence enable an easier real-time DSP implementation for industrial applications.

5.2. Neuro-fuzzy inference

5.2.1. ANFIS model structure

The previous two simple classification methods may not work reliably when the data patterns become more complicated. Using the statistics of the wavelet components as raw features and a neural network for classification should provide a more robust diagnostic method. Through training the neural network the diagnostic system should be adaptive to minor changes that can cause variations in the response to each pulse [18]. The adaptive neural fuzzy inference system (ANFIS) was thus used to learn information about the three patterns.

Jang first introduced the adaptive network-based fuzzy inference system (ANFIS) in 1993 [24]. It is a model that maps inputs through input membership functions (MFs) and associated parameters, and then through output MFs to outputs. The initial membership functions and rules for the fuzzy inference system can be designed by employing human expertise about the target system to be modelled. ANFIS can then refine the fuzzy if–then rules and membership functions to describe the input-output behaviour of a complex system. Jang showed that even if human expertise is not available it is possible to intuitively set-up reasonable membership functions and employs the neural training process to generate a set of fuzzy if-then rules that approximate a desired data set [17,24].

Fig. 7 shows the structure of an ANFIS with two inputs, four rules and one output. The input membership function layer performs fuzzification of the inputs. For each input, a fuzzy set A in X

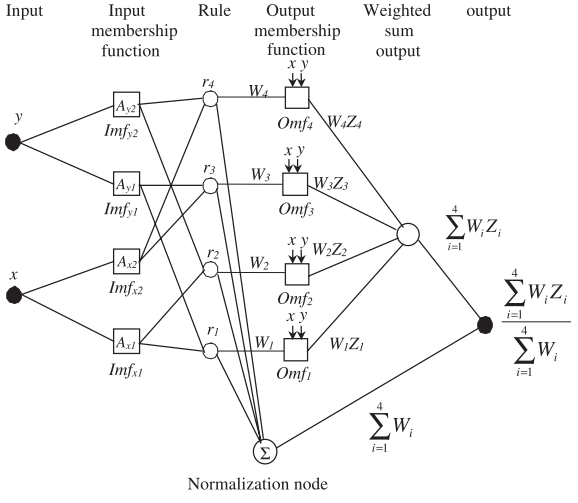

Fig. 7. ANFIS model structure.

(the universe of disclosure) is defined as set of ordered pairs:

$$A = \{x, \mu_A(x) | x \in X\}.$$

$\mu_A(x)$ is called the membership function of x in A, which maps each element of X to a membership value between 0 and 1. For example, the input membership function node Imf_{x1} takes the input value x, performs fuzzification mapping using the membership function defined in fuzzy set A_{x1} and then outputs a fuzzy number $\mu_{Ax1}(x)$. Commonly used membership functions include piecewise linear functions, the Gaussian distribution function, the sigmoid curve, quadratic and cubic polynomial curves, etc.

The rule layer applies fuzzy operators (*AND*, *OR*, *NOT*) to the antecedent and resolves the antecedent to a new fuzzy number, which is a degree of support for the rule. In this paper, the fuzzy operator *AND* is used; for example, the first rule defined for the ANFIS model in Fig. 7 is:

Rule 1. IF (x is in A_{x1} *AND* y is in A_{y1}) THEN (Output is Omf_1) (weight = 1)

The *AND* operation can be either product or minimum, for example, for the above rule,

$$\text{Product}: \quad W_1 = \mu_{Ax1}(x) \times \mu_{Ay1}(y),$$

$$\text{Minimum}: \quad W_1 = \min\{\mu_{Ax1}(x), \mu_{Ay1}(y)\}.$$

In the Sugeno-type fuzzy inference system, the output membership functions are usually a constant ($Z_i = c_i$) or a linear function ($Z_i = p_i x + q_i y + c_i, p_i$ and q_i are parameters introduced to the adaptive nodes in the output membership function layer). Higher than second order output membership functions can introduce significant complexity and thereby slow down the training without obvious merits in performance [17,24,25].

Then, in the weighted sum output layer, the weighted output of the consequence parameters are summed up ($\Sigma W_i Z_i$). The normalised node calculates the sum of the weight functions for all the rules (ΣW_i), and finally, the output node computes the normalised weighted output ($\Sigma W_i Z_i / \Sigma W_i$).

When fuzzy inference is applied to a system for which a collection of input/output data is available for modelling, the parameters associated with the membership functions could be selected so as to tailor the membership functions to the input/output data in order to account for specific types of variations in the data values being used. This is where the so-called neuro-adaptive learning techniques incorporated in fuzzy inference are useful. The parameter tuning, or what is known as learning in neural network terminology, can be performed using either a back propagation or least squares method [24,25].

5.2.2. Implementation of ANFIS for diagnostic classification

As a first computational experiment, the inference system for the bearing fault classification problem is constructed as a Sugeno-type inference system with six inputs (the six elements of the feature vector) and one output (the decision variable). For training the targets are coded as: 1 (normal), 0 (inner race fault) and –1 (ball fault). Each of the output membership functions is simply a constant (zero-order Sugeno MF). For each input as well as the output, two membership functions are defined and a pi-shaped non-linear function (Fig. 8(a)) is selected arbitrarily. Each rule is assigned a unit weight.

To train the model, the training data was prepared using the data in Tables 2–4. The training quickly converged and terminated at the tenth epoch with a training accuracy (average error) of

2.04126×10^{-7}. The testing error is at the same level (10^{-7}). Fig. 9(a) shows the comparison between the real data and the predicted output using the trained model; Fig. 9(b) shows the prediction error. If a linear membership function is used for the output layer, the number of linear

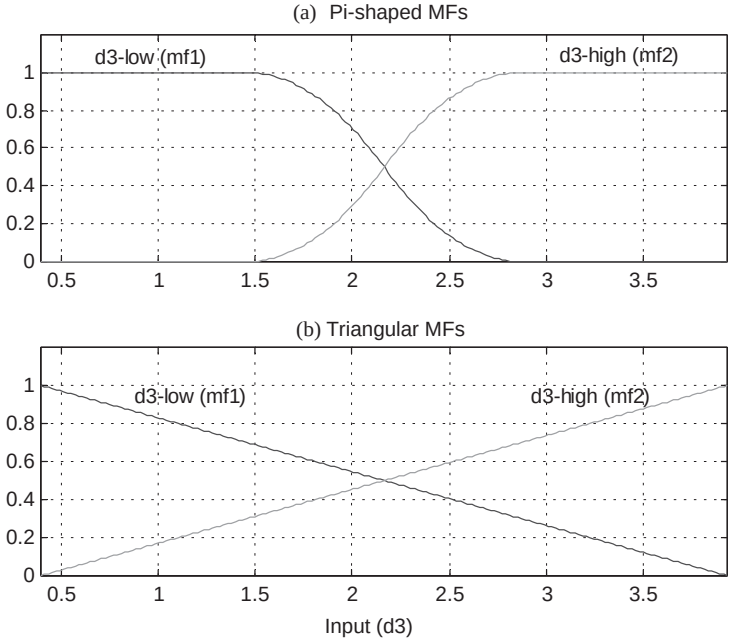

Fig. 8. Input membership functions used in the ANFIS.

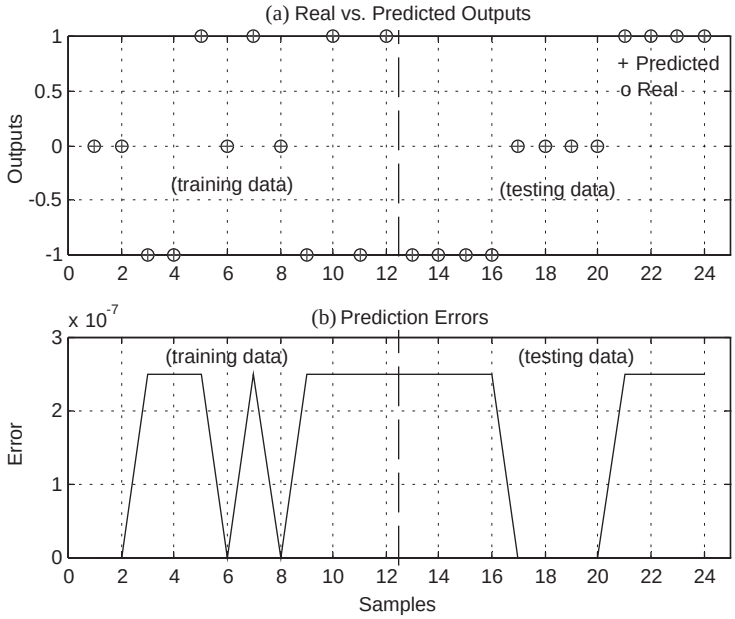

Fig. 9. Prediction results using the trained ANFIS model (6 inputs).

parameters will increase dramatically to 448, which greatly slows down the training, and at the same time increases the training error to 6.39243×10^{-6}. When training an ANFIS, it is often easiest to use linear input membership functions. However, in situations where there is no knowledge available about the linear separability of the pattern association problem, non-linear membership functions should first be tried for the ANFIS model. Using a linear model for a non-linear problem can lead to significant performance degradation. In our application, when the pi-shape input membership functions are replaced with triangle functions (Fig. 8(b)), the training error increases to 10^{-4}, and the testing error increases to the order of 10^{-2}.

On the other hand, regarding the performance issue, the small training error reminds us of a potential over-training problem in ANFIS learning. The number of rules in the above ANFIS structure may be too large for the small set of training data available. This results in a large number of nodes and interconnections since there will be 64 nodes in both 'rule' and 'outputmf' (output membership function) layers in the network structure as depicted in Fig. 7. To guarantee the generalisation capability of the trained network, a large training data set and extensive training efforts should be required. As a consequence, the decision surfaces that result from extensive training can be quite complex. When the size of the training set is not large enough, this can lead to a situation in which the network merely becomes tuned to the particular training set, rather than adjusting itself to recognise all members of the classes at large. One can envision a partitioning of the feature space wherein the network has placed a small hyperellipsoid around each point in a small training set. This will, of course, produce a low error on the training set, but poor performance in general [21,24].

For the data set available, to use the ANFIS appropriately, we resort to using the most superior features as the inputs to a simpler ANFIS. By looking at the plots in Figs. 5–6, we notice that the approximation a_5 and details d_3 and d_4 should be superior to the others in classifying the three classes. Further analysis shows that d_3 and d_4 are linearly correlated. Therefore, a_5 and d_3 (or d_4) were selected as the two superior features. For more general problems of feature superiority checking, a statistical index called *Class Separation Distance* [21] can be used; to perform dimension reduction of feature vectors, *principle component analysis* (PCA)[21] *or* '*measure of significance*'[18] can also be used.

With two input nodes (with two pi-shaped membership functions for each input) and an output node (with constant membership functions), the new ANFIS has the same structure as shown in Fig. 7 (with four rules and four weighted-sum-output nodes). The training quickly converges at the second epoch with a training accuracy of 0.00997871. Fig. 10 shows the computational results when a_5 and d_3 were used as the two inputs to the ANFIS. If features other than a_5 with d_3 or d_4 are used, mis-classification may even occur, which verifies the significance of these salient features. Fig. 11 shows the computational result when d_2 and d_5 were used as the two inputs to the ANFIS.

6. Discussions and conclusions

A new scheme has been developed for the diagnosis of defects in ball bearings. The technique is based on statistical analysis, the discrete wavelet transform, and pattern classification techniques such as neuro-fuzzy inference. By using vibration data collected from an AC motor driven system with different faulted bearings installed, this diagnostic strategy was evaluated. The signals were

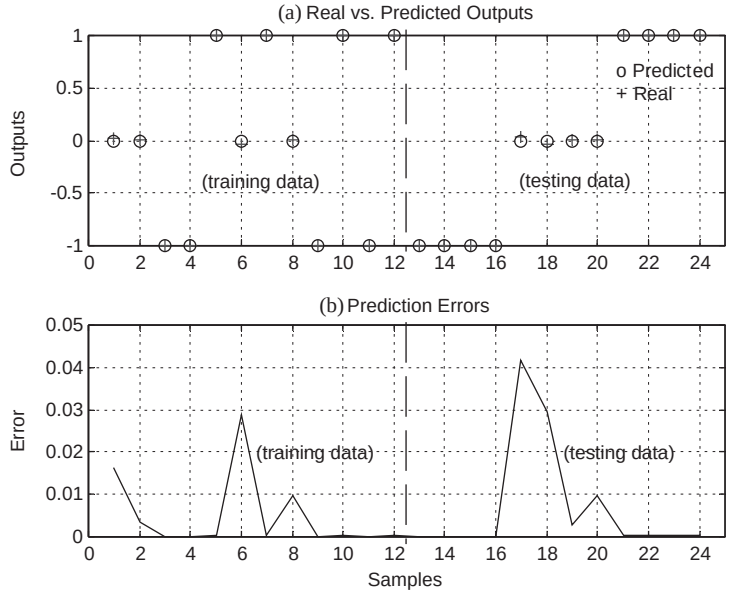

Fig. 10. Prediction results using the trained ANFIS model (d_3 and a_5 as inputs).

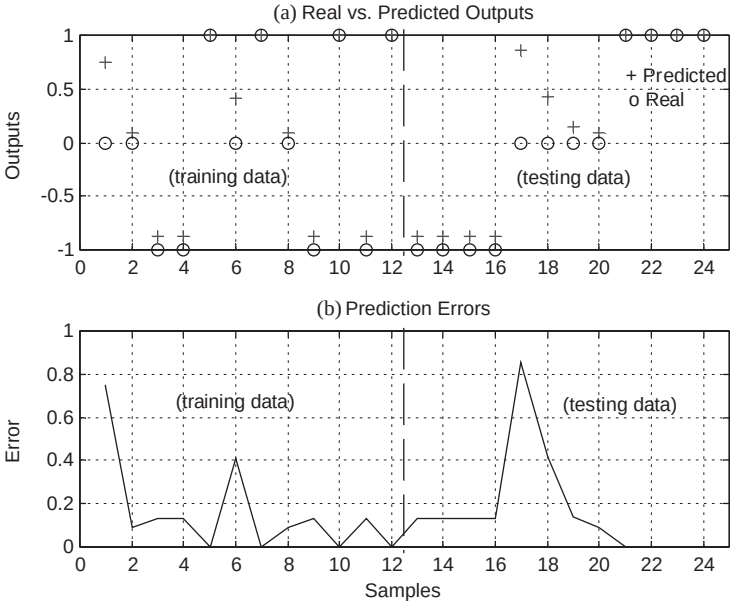

Fig. 11. Prediction results using the trained ANFIS model (d_2 and d_5 as inputs).

normalised to (0,1) standard random variables, and then the wavelet transforms were performed using the Daubechies-2 wavelet. Feature vectors were first formed by using all the component as the vector elements. In the decision making stage, an ANFIS was trained as the pattern classifier. For comparison purposes, the Euclidean vector distance method as well as the vector correlation

coefficient method were also investigated. Both vector distance and correlation coefficient techniques are simple, and easy to implement, and the former works better than the latter in this example. The performance of the ANFIS learning was also addressed and two salient features were selected for ANFIS training. The results also show that ANFIS is a good candidate for future development work because of its non-linear approximation capability and adaptability. It is expected that using wavelet analysis and fuzzy math, it may be possible to identify the time of occurrence and the degree of severity of the fault—a first step toward prognostics.

To further investigate on incipient fault detection and fault growth monitoring, additional experiments have been designed and vibration data has been collected for bearings with different sizes of faults on the races and the rolling elements. The fault sizes were designed to be 7, 14 and 21 mils respectively, which are much smaller than the 40 mils used in the previous analysis. The Daubechies-10 wavelet [22,25] was used to perform the transforms. By analysing the approximations and the different levels of detail, it was found that some characteristic components associated with the continual increase in fault severity (fault size from 7 to 21 mils) as well as abrupt changes in defect size could be detected. Using the wavelet transform together with fuzzy logic to quantify the degree of severity of an incipient fault is a promising technique for prognostics. For detailed discussions, please refer to [19]. Further investigations should be conducted on optimal wavelet decomposition in the sense of best performance in incipient fault detection, isolation and severity monitoring. A more challenging task is to explore identifying simultaneous multiple faults through the smart use of time-scale analysis and other techniques in systems science and engineering.

Acknowledgements

This work was supported in part by the Office of Naval Research under agreement N00014-98-3-0012 and the National Science Foundation, Grant ECS-9906218.

References

[1] R.R. Schoen, T.G. Habetler, F. Kamran, R.G. Bartheld, Motor bearing damage detection using stator current monitoring, IEEE Transactions on Industrial Applications 31 (6) (1995) 1274–1279.

[2] T.B. Brotherton, T. Pollard, Applications of time-frequency and time-scale representations to fault detection and classification, Proceedings of the IEEE Signal Processing International Symposium on Time-Frequency and Time-Scale Analysis, Orlando, FL, 1992, Vol. 2242, pp. 95–98.

[3] D.–M. Yang, A.F. Stronach, P. MacConnel, Third-order spectral techniques for the diagnosis of motor bearing condition using artificial neural network, Mechanical Systems and Signal Processing 16 (2-3) (2002) 391–411.

[4] J.A. Leonard, M.A. Kramer, Radial basis function networks for classifying process faults, IEEE Control Systems Magazine (1991) 31–38.

[5] M. Chow, R.N. Sharpe, J. Hung, On the application and design of artificial neural network for motor fault detection—II, IEEE Transactions on Industrial Electronics 40 (2) (1993) 189–196.

[6] L.P. Heck, K.C. Chou, Gaussian mixture model classifier for machine monitoring, Proceedings of the IEEE world Congress on Computational Neural Network and International Conference on Intelligence, Vol. 7, 1994, pp. 4493–4496.

[7] E. Meyer, T. Tuthill, Bayesian classification of ultrasound signals using wavelet coefficients, Proceedings of the IEEE National Aerospace and Electronics Conference (NAECON), Vol. 1, 1995, pp. 240–243.
[8] K.W. Baugh, On Parametrically Phase-Coupled Random Harmonic Processes, Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statics, 1993, pp. 346–350.
[9] H.C. Choe, C.E. Poole, A.M. Yu, H.H. Szu, Novel Identification of Intercepted Signals from Unknown Radio Transmitters, Proceedings of the SPIE Wavelet Applications 2419 (1995) 504–517.
[10] T. Peng, W. Gui, M. Wu, Y. Xie, A fusion diagnosis approach to bearing faults, Proceedings of the International Conference on Modeling and Simulation in Distributed Applications, 2001, pp. 759–766.
[11] H.C. Choe, R.E. Karlsen, G.R. Gerhart, T.J. Meitzler, Wavelet-based ground vehicle recognition using acoustic signals (invited paper), Proceedings of SPIE Wavelet Applications 2762 (1996) 434–445.
[12] M. Desai, D.J. Shazeer, Acoustic Transient Analysis Using Wavelet Decomposition, Proceedings of the IEEE Conference on Neural Networks for Ocean Engineering, 1991, pp. 29–40.
[13] L.P. Heck, K.C. Zhou, Feature extraction based on minimum classification error/generalized probabilistic descent method, Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing 6 (1994) 133–136.
[14] B.R. Bakshi, A. Koulouris, G. Stephanopoulos, Wave-net: novel learning techniques, and the indication of physically interpretable models, Proceedings of the SPIE Wavelet Applications (1994) 637–648.
[15] S. Kadambe, Text independent speaker identification system based on adaptive wavelets, Proceedings of the SPIE Wavelet Applications 2242 (1994) 669–677.
[16] B. Liu, S.F. Ling, On the selection of informative wavelet for machinery diagnosis, Mechanical Systems and Signal Processing 11 (3) (1999) 145–162.
[17] J. Altmann, J. Mathew, Multiple band-pass autoregressive demodulation for rolling-element bearing fault diagnosis, Mechanical Systems and Signal Processing 15 (5) (2001) 963–977.
[18] P. Xu, A.K. Chan, Fast and robust neural network based wheel bearing fault detection with optimal wavelet features, Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN'02), Vol. 3, 2002, pp. 2076–2080.
[19] X. Lou, Fault detection and diagnosis for rolling element bearing, Ph.D. thesis, Case Western Reserve University, 2000.
[20] R.S. Liptser, A.N. Shiryaev, Statistics of Random Processes, Springer-Verlag, New York, 1978.
[21] K.R. Castleman, Digital Image Processing, Prentice-Hall, Englewood Cliffs, NJ, 1998.
[22] I. Daubechies, Ten lectures on wavelets, CBMS-NSF Series in Applied Mathematics (SIAM), 1991.
[23] Mathworks, Wavelet Toolbox—for Use with MATLAB®, 1998 User manual of Mathworks.
[24] J.-S.R. Jang, ANFIS: adaptive-network-based fuzzy inference systems, IEEE Transactions on Systems, Man, and Cybernetics 23 (3) (1993) 665–685.
[25] Mathworks, Fuzzy Logic Toolbox—for Use with MATLAB®, User manual of Mathworks, 2000.

Further reading

I.J. Booth, K.H.V. Booth, Using neural nets to identify marine mammals, Proceedings of the IEEE OCEANS'93 3 (1993) 112–115.
G. Lundberg, A. Palmgren, Dynamic capacity of rolling bearings, Acta Polytechnica 96 Mechanical Engineering Series 2, 1952.
N.G. Nikolaou, I.A. Antoniadis, Demodulation of vibration signals generated by defects in rolling element bearings using complex shifted Morlet wavelets, Mechanical Systems and Signal Processing 16 (4) (2002) 677–694.