Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Sound and Vibration 277 (2004) 1005–1024

JOURNAL OF SOUND AND VIBRATION

www.elsevier.com/locate/jsvi

Machine fault diagnosis through an effective exact wavelet analysis

Peter W. Tse[a,]*, Wen-xian Yang[b], H.Y. Tam[a]

[a] *Smart Asset Management Laboratory, Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Tat Chee Ave., Kowloon, Hong Kong, China*

[b] *Institute of Vibration Engineering, Northwestern Polytechnic University, Xi'an, 710072, Shaanxi, China*

Received 26 July 2002; accepted 25 September 2003

Abstract

Continuous wavelet transforms (CWTs) are widely recognized as effective tools for vibration-based machine fault diagnosis, as CWTs can detect both stationary and transitory signals. However, due to the problem of overlapping, a large amount of redundant information exists in the results that are generated by CWTs. The appearance of overlapping can smear the spectral features and make the results very difficult to interpret for machine operators. Misinterpretation of results may lead to false alarms or failures to detect anomalous signals. Moreover, as conventional CWTs only use a single mother wavelet to generate daughter wavelets, the distortion of the original signal in the resultant coefficients is inevitable. Obviously, this will significantly affect the accuracy in anomalous signal detection. To minimize the effect of overlapping and to enhance the accuracy of fault detection, a novel wavelet transform, which is named as *exact wavelet analysis*, has been designed for use in vibration-based machine fault diagnosis. The design of exact wavelet analysis is based on genetic algorithms. At each selected time frame, the algorithms will generate an adaptive daughter wavelet to match the inspected signal as *exactly* as possible. The optimization process of *exact wavelet analysis* is different from other adaptive wavelets as it considers both the optimization of wavelet coefficients and the satisfaction of the admissibility conditions of wavelets. The results obtained from simulated and practical experiments prove that exact wavelet analysis not only minimizes the undesirable effect of overlapping, but also helps operators to detect faults and distinguish the causes of faults. With the help from exact wavelet analysis, sudden shutdowns of production and services due to the fatal breakdown of machines could be avoided.

*Corresponding author.

E-mail address: meptse@cityu.edu.hk (P.W. Tse).

0022-460X/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsv.2003.09.031

1. Introduction

1.1. The use of wavelet transforms in vibration-based machine fault diagnosis

The Fourier transform (FT) represents a signal by a family of complex exponents with infinite time duration. Therefore, FT is useful in identifying harmonic signals. However, due to its constant time and frequency resolutions, it is weak in analyzing transitory signals. In contrast, CWTs have a constant frequency to bandwidth ratio analysis. Therefore, CWTs provide powerful multi-resolution in time–frequency analysis for characterizing the transitory features of non-stationary signals. Moreover, CWTs can decompose an inspected signal into a family of elementary functions. This ability renders the analysis of the inspected signal easier for machine operators. Hence, extensive research has been conducted on the use of CWTs in vibration-based machine fault diagnosis [1–4]. The authors had also performed a comparison on the effectiveness of the popular envelope detection and CWTs on the fault diagnosis of roller bearings [5]. The study shows that CWTs outperform the envelope detection in identifying the causes of faults. Nevertheless, CWTs have not yet been widely adopted by industry for machine fault diagnosis. The major obstacle is the difficulty in interpreting the results that are generated by CWTs. Fig. 1 shows the results that were generated by the Morlet CWT for analyzing a vibration signal that was collected from a defective roller bearing.

1.2. The deficiency of conventional wavelet transforms

Every component has multiple excited frequency zones, particularly in the high-frequency region. The reason to seek for the excited frequency zone is to obtain higher signal-to-noise ratio

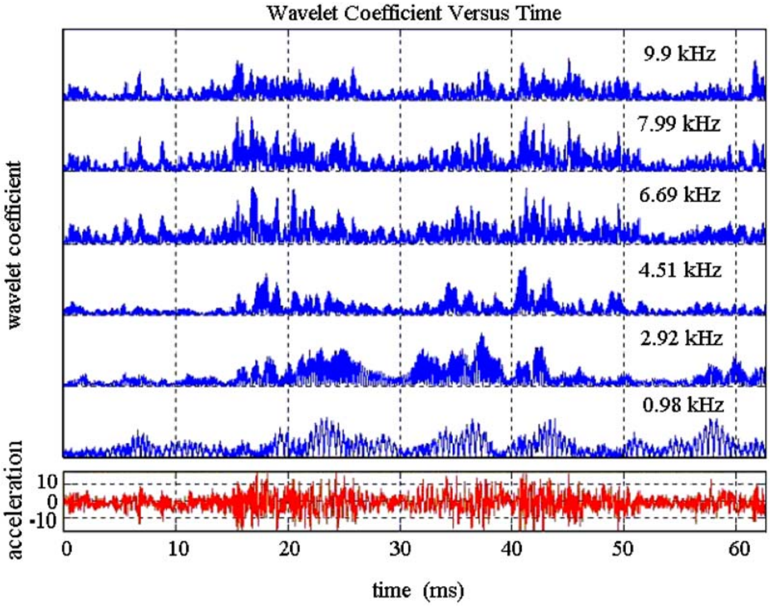

Fig. 1. Overlapping occurs in the result that is generated by the Morlet CWT.

so that fault related impulses or impacts could be revealed more clearly. Normally, if the component is defective, then the fault-related impulses, which are caused by the contacts between the surface of the defective component and other surfaces during rotation, should be revealed more clearly in the excited frequency zones [6]. The excited frequency zones of the defective roller bearing were found by performing an impact test on the bearing housing. The theories of the impact test can be found in a number of publications by Brual and Kjaer, such as the one from Randall [7]. Due to the limitation of accelerometers, we recorded the vibration generated by the bearing housing up to 10 kHz. When performing the impact test on the defective roller bearing, it was found that the one of the excited frequency zones of the bearing's housing was around 8 kHz. In Fig. 1, high values of wavelet coefficients, which are related to faulty impulses, are found at the frequency level of 7.99 kHz. This frequency level coincides well with the 8 kHz excitation frequency found by the impact test. Hence, the identification of excited frequency zones for the inspected component helps the operator to reveal problems of the component. Unfortunately, high values of coefficients also emerge at the same time in adjacent frequency levels of 6.69 and 9.99 kHz. These undesirable high coefficients are caused by *overlapping*. Although their values are relatively smaller than the ones at 7.99 kHz, they smear the spectral features and make the result very difficult to be interpreted by machine operators. As each component of a running machine has its own excited frequency zones, the appearance of high coefficients at adjacent frequency levels may mislead operators to think that faults are also occurring in other components simultaneously. Consequently, the overlapping induces operators to misinterpret the results and make incorrect decisions in fault diagnosis. Misinterpretation of results may lead to false alarms or the failure to detect anomalous signals. Such negligence may cause fatal breakdown of machines, which could interrupt production and services. In the worst scenario, it could even cause human casualties.

Besides the problem of overlapping, the results that are generated by conventional CWTs always contain distortions as compared to the original inspected signal. The Morlet CWT has been selected to demonstrate the cause of distortions as it is often used in decomposing vibration signals for machine fault diagnosis. As shown in Fig. 2, the inspected signal contains three transitory features—two harmonic waveforms and one impulsive signal. The coefficients that are generated by the Morlet CWT at scales X, Y and Z clearly show that the three transitory features are diluted and distorted at all three scales. Hence, the original time and frequency properties of the inspected signal are difficult to identify from the displayed coefficients in these scales. This undesirable effect makes the identification of anomalous signals even more difficult. The distortions as shown in Fig. 2 are mainly due to the fact that Morlet CWT uses only one single mother wavelet to generate the required family of wavelets. Obviously, a single mother wavelet

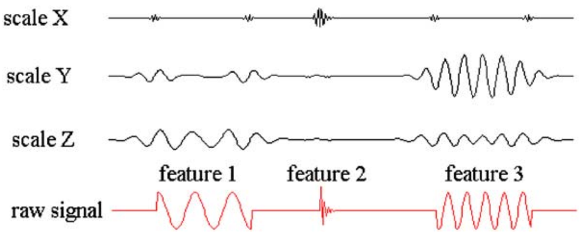

Fig. 2. Features being diluted and distorted in the analysis of Morlet CWTs.

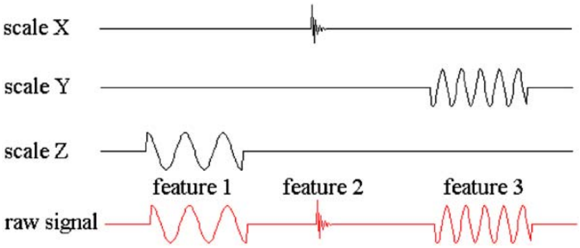

Fig. 3. The expected ideal result of *exact analysis*.

cannot provide all necessary wavelets to adaptively match with each of the characteristics of the transitory features that are contained in the raw signal [8].

Idealistically, each of the three temporal features of the raw signal should only appear in one scale that has the same frequency content and defined time frame as shown in Fig. 3. However, due to the problems of overlapping in adjacent scales and the distortion of the signal, the three temporal features appear in all three scales of the defined time frame as shown in Fig. 2. For the benefit of vibration-based machine fault diagnosis, the decomposed features that are obtained from an effective analyzing tool should possess all of the information on the amplitude, time and frequency *exactly* as they are in the original raw signal. That is, after the decomposition of the raw signal, each temporal feature should appear only in its expected scale and time frame *exactly* as it is displayed in the raw signal. The ideal results should have no overlapping, distortion or redundant information. Such a clear and precise result is called *exact analysis*. The aim of this paper is to develop an effective algorithm that will allow CWTs to achieve such a desirable result.

1.3. The limitations of current methods used for minimizing overlapping

Wavelet ridge extraction [9] has been used to minimize the redundant information that often appears in the results generated by conventional CWTs. However, it cannot solve the problem as precisely as that performed by exact wavelet analysis. One of the main reasons for this deficiency is that it uses a single mother wavelet to generate the required family of wavelets. Another reason is that although it has the ability to delete some disturbance information, the deleted information may not include the undesired information which is caused by overlapping. The method of matching pursuit [10] uses a sub-optimal set of atoms that are selected from a time–frequency dictionary to solve the overlapping problem. However, this method still cannot generate the same results as exact wavelet analysis because it uses a single function to produce the time–frequency atoms. Many methods had also been tried to obtain the best time–frequency resolutions in order to minimize the effects of distortion and dilution to the original signal. For instance, Telfer et al. [11] optimized the shift and dilation parameters of the discretization of a chosen wavelet transform. Szu et al. [12] sought the optimal linear combination of predefined wavelet bases for the classification of speech signals. Kocur et al. [13] employed a neural network to select wavelet features for breast cancer diagnosis. Tagliarini and Page [14] optimized the wavelet coefficients so that they could provide desirable properties for image identification. Galvao et al. [15] calculated adaptive biased wavelet expansions by using the conventional gradient-descent method. Finally, Silva [16] studied evolutionary-based methods for adaptive signal representation. These adaptive methods are superior in feature extraction than the methods that use predefined wavelets (e.g., the Daubechies family). To summarize the achievements of these studies, they have adopted either one of two optimizing strategies for constructing adaptive wavelets. The first strategy is the optimization of the scale and translation factors for a single and predefined mother wavelet. The second strategy is the direct optimization of the wavelet coefficients. For the first strategy, in spite of the efforts that have been spent on optimizing the scale and the translation factors, a single form of mother wavelet cannot possibly generate all of the required daughter wavelets to exactly match all of the transitory features of an inspected signal. The second strategy cannot accomplish the similar result as produced by exact wavelet analysis because of the admissibility conditions of wavelet [17] have been neglected during the process of optimization. This negligence directly decreases the matching ability and the accuracy of the analyzed results. To overcome the above limitations, exact wavelet analysis has been designed to produce an exact analysis for any inspected signal. It is aimed at compensating the inherent deficiency of conventional CWTs and minimizing the undesirable effects that often occur in their generated results for vibration-based machine fault diagnosis.

1.4. Introduction on exact wavelet analysis

Two versions of exact wavelet analysis have been developed by the authors to produce an exact analysis for the inspected signals. The first version utilizes the concept of 'maximum matching mechanism' to determine the most appropriate coefficients to represent the inspected raw signal. A common phenomenon has been observed when using the conventional CWTs to decompose a given signal. Within the selected time frame, if a daughter wavelet, which is generated by a particular scale, has the largest value of wavelet coefficient, it often implies that the shape of that daughter wavelet can match the shape of the inspected signal better than other daughter wavelets generated by other scales. Such a phenomenon is the so-called 'maximum matching mechanism' [18,19]. The advantage of the first version implemented by such concept is simple and fast in computation. Its disadvantage is that it cannot guarantee that the selected daughter wavelet will have a geometric shape exactly similar to the inspected signal within the selected time frame. The major problem is the selection of mother wavelet is not adaptive to the inspected signal. Within each selected time frame, this method only provides a relative measure to determine which scale can generate a daughter wavelet to match for the inspected signal better than other scales of the same mother wavelet. The first version cannot achieve the desirable exact analysis for the inspected signal.

In view of this, a second version of exact wavelet analysis has been designed and implemented. The second version is aimed to provide a direct measure of the similarity in shapes between the daughter wavelet and the inspected signal. Instead of using the largest value of wavelet coefficient, the 'normalized dot product' of the daughter wavelet and the inspected signal is adopted for measuring their similarity in shape. Genetic algorithms are employed to optimize not only the scale and translation factors, but also the formation of mother wavelets that are used to generate a series of daughter wavelets. Therefore, the derived exact wavelets are able to match all of the properties of the inspected signal as closely as possible. Although the computation is more intensive, the results are more precise and easier for machine operators to identify anomalous vibrations generated by defective machines.

The paper is organized as follows. The cause and effect of overlapping that often occurs in the results generated by conventional CWTs are investigated and discussed in Section 2. The algorithm and the design of exact wavelet analysis are introduced in Section 3, and a verification of the effectiveness of the analysis is followed. Section 4 describes the set up of an industrial test that is used to further verify the effectiveness of exact wavelet analysis in practice. Finally, Section 5 provides a discussion of the results and benefits of using exact wavelet analysis in vibration-based machine fault diagnosis, and draws conclusions.

2. The cause and effect of overlapping

The equation of CWTs that are applied to an inspected signal $x(t)$ can be expressed as

$$CWT_x(a,b) = <\psi_{a,b}(t)\cdot x(t)> = |a|^{-1/2}\int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)\,\mathrm{d}t, \tag{1}$$

where $\langle\cdot\rangle$ indicates the inner product; the asterisk, '$*$', stands for complex conjugate; and $\Psi_{a,b}(t)$ denotes the daughter wavelets that are derived from the mother wavelet $\Psi(t)$ by continuously varying both the scale factor a, and the translation or time shift factor b. The factor $|a|^{-1/2}$ is used to ensure energy conservation.

Eq. (1) states that the generated daughter wavelets are dependent on the variation of the translation b and the scale a. It manifests the signal $x(t)$ at different levels of resolution by measuring the similarity of signal $x(t)$ and the daughter wavelet $\Psi_{a,b}(t)$ at different scales. This implies that, if the shape of a particular daughter wavelets at scale a' is closely matched with the shape of a segmental signal $x(t)$ at time shift factor b', then the wavelet-transforming coefficient $CWT_x(a',b')$ will have a large magnitude. However, this scenario does not imply that the magnitude of coefficient $CWT_x(a'',b')$ is zero even when the shape of another daughter wavelet at arbitrary scale a'' deviates from the shape of the segmental signal. Consequently, at time shift factor b', the overlapping of coefficients occurs in all adjacent scales. The closer the scale a'' approximates to a', the more serious the overlapping may occur. Hence, as long as CWTs are used to decompose the inspected signal for identifying the properties of the signal, the undesirable effect of overlapping is inevitable. Fig. 4 illustrates the effect of overlapping by showing part of the coefficients that are generated from the Morlet CWT to match for a simulated signal $x(t)$ which consists of two temporal sinusoidal waveforms as

$$x(t) = x_1(t) + x_2(t) \quad (0.05 \leqslant t \leqslant 0.2), \tag{2}$$

where

$$x_1(t) = \begin{cases} 10\sin(400\pi t) & 0.075 < t < 0.1\ \mathrm{s} \\ 0 & \text{otherwise} \end{cases},$$

and

$$x_2(t) = \begin{cases} 10\sin(200\pi t) & 0.125 < t < 0.15\ \mathrm{s} \\ 0 & \text{otherwise} \end{cases}.$$

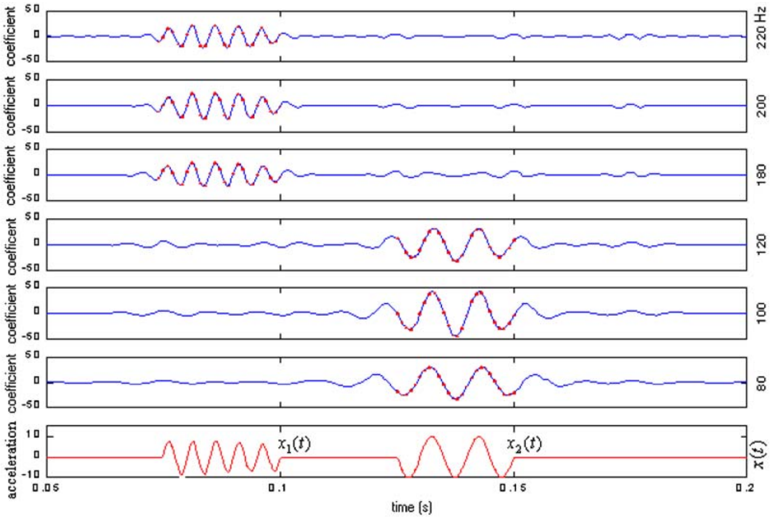

Fig. 4. Overlapping occurs in the adjacent scales of the simulated signal $x(t)$.

According to the definition of Eq. (2), the results of the Morlet CWT should only exist as a temporal waveform $x_1(t)$ from 0.075 to 0.1 s at frequency level 200 Hz, and another temporal waveform $x_2(t)$ from 0.125 to 0.15 s at frequency level 100 Hz. However, as shown in Fig. 4, the unexpected waveforms appear also at levels of 220 and 180 Hz, which are adjacent to 200 Hz. These waveforms are redundant and caused by overlapping. A similar phenomenon also occurs at levels of 120 and 80 Hz, which are adjacent to 100 Hz. The problem of overlapping exists in the results generated not only by the Morlet CWT, but also by other conventional types of CWTs. Thus, this undesirable effect of overlapping must be minimized to ensure an accurate and reliable decomposition of vibration signals for machine fault diagnosis.

3. The design of exact wavelet analysis

3.1. Satisfaction of the admissibility conditions

To minimize the effects of overlapping and distortion, the derived exact wavelets are rendered to be adaptive to the inspected signal via optimization. In the process of optimizing the derived exact wavelets, two conditions must be satisfied. The first condition is to ensure the derived wavelets satisfy the admissibility conditions. The second condition is to identify the required variables of the exact wavelets to be optimized. The '*abc* wavelets' that were proposed by Borde [20] made a good solution for the first condition. The distinguishing feature of *abc* wavelets is their extra parameter *c*. This extra parameter offers another degree of freedom to wavelet function so that it can mimic any signal better than the conventional '*ab*' two-dimensional wavelets. The *abc* wavelets are derived by solving a system of equations that are specially defined for describing the admissibility conditions of wavelets [21]. The equations required ensuring a 4-tap wavelet to satisfy the admissibility conditions are:

$$\begin{array}{ll} c_0^2+c_1^2+c_2^2+c_3^2=1 & \text{orthonormality,}\\ c_0c_2+c_1c_3=0 & \text{orthogonality,}\\ c_3-\alpha c_2+\beta c_1-\eta c_0=0 & \text{lock condition,}\\ c_0+c_1+c_2+c_3=\sqrt{2} & \text{energy conservation.} \end{array} \tag{3}$$

In Eq. (3), each individual equation describes an admissibility condition of the 4-tap wavelet. The variables c_i ($i=0,1,2,3$) define the four taps of the wavelet. The value $\sqrt{2}$ in the fourth equation is used to ensure energy conservation. The parameters α, β, and η can be obtained through an arbitrarily designed continuous function $f(\vartheta,\xi)$, where

$$\begin{aligned} \alpha &= f(\vartheta=1,\xi),\\ \beta &= f(\vartheta=2,\xi),\\ \eta &= f(\vartheta=3,\xi). \end{aligned} \tag{4}$$

By solving the equations that are listed in Eq. (3), the taps of the 4-tap wavelet can be derived as

$$\begin{aligned} c_0 &= \frac{1-\alpha+2\alpha^2-2\beta-\alpha\beta+\beta^2+\eta-2\alpha\eta+\beta\eta}{2\sqrt{2}[(1-\beta)^2+(\eta-\alpha)^2]}\pm(1-\beta)\gamma,\\ c_1 &= \frac{2-\alpha+\alpha^2-2\beta+\alpha\beta-\eta-2\alpha\eta+\beta\eta+\eta^2}{2\sqrt{2}[(1-\beta)^2+(\eta-\alpha)^2]}\pm(\alpha-\eta)\gamma,\\ c_2 &= \frac{1+\alpha-2\beta+\alpha\beta+\beta^2-\eta-2\alpha\eta-\beta\eta+2\eta^2}{2\sqrt{2}[(1-\beta)^2+(\eta-\alpha)^2]}\pm(\beta-1)\gamma,\\ c_3 &= \frac{\alpha+\alpha^2-2\beta-\alpha\beta+2\beta^2+\eta-2\alpha\eta-\beta\eta+\eta^2}{2\sqrt{2}[(1-\beta)^2+(\eta-\alpha)^2]}\pm(\eta-\alpha)\gamma, \end{aligned} \tag{5}$$

where

$$\gamma = \frac{\sqrt{1+2\alpha+\alpha^2-6\beta+2\alpha\beta+\beta^2+2\eta-6\alpha\eta+2\beta\eta+\eta^2}}{2\sqrt{2}[(1-\beta)^2+(\eta-\alpha)^2]}. \tag{6}$$

By using the above generic solutions, one can easily derive various scales of wavelets that satisfy the admissibility conditions.

3.2. Identifying the required variables to be optimized

To ensure that the derived wavelets may closely match the inspected signal, one must identify the required parameters of the exact wavelets to be optimized. We have modified the simple Borde exponent function [21] to identify the required parameters. The number of variables used in the function is increased from two variables to four variables as

$$f(\vartheta,\varsigma,\Omega,\theta)=\mathrm{e}^{-\vartheta^2\varsigma}\sin(\Omega\varsigma+\theta), \tag{7}$$

where $\vartheta = 1, 2$ and 3 (as defined in Eq. (4)), ς and Ω are two positive real numbers, and the phase angle θ varies from 0 to 2π. Note that the function $f(\vartheta, \varsigma, \Omega, \theta)$ uses four variables, which provide more degrees of freedom to the derived wavelets to match the properties of the inspected signal.

To optimize the selection of the four variables, a comprehensive optimization strategy has been adopted for the construction of the exact wavelets. As mentioned in Section 1, both the wavelet coefficients and the scale factor of the exact wavelets should be optimized simultaneously. Note that the variables ς, Ω and θ, and the scale a will be optimized during the optimization process. The variable ϑ in the function $f(\vartheta, \varsigma, \Omega, \theta)$ will not be optimized as its value has already been defined in Eq. (4). The translation or time shift factor b in Eq. (1) is also not required to be optimized because it is defined as the shift of one unit of time, the smallest possible time shifting unit equivalent to the sampling time. The inspected signal will be optimized point by point per unit of time until all the data points contained in the signal have been optimized.

3.3. Selection of the optimizing method and the fitness function

When selecting an appropriate method for optimization, one must consider the trade-off between the intensiveness of computation and the accuracy of the results. Many optimization methods are available, and each has its own advantages and limitations. The conventional gradient-descent and conjugated gradient [22] methods can lead to local minima and maxima. The use of the flexible polyhedron [23] involves intensive computation. Neural networks can be used to for global optimization [24]. However, it is well known that the architecture of neural networks greatly affects their performance and accuracy. The selection of an optimal architecture for a neural network is a tedious and computationally intensive task. Another alternative is the use of genetic algorithms, which have the ability and confirmed performance in global optimization [25], as long as the algorithm used is not too computationally intense. Fortunately, most industrial machines do not require real-time and on-line fault diagnosis. Therefore, genetic algorithms have been employed to optimize the variables for deriving the required exact wavelets that are adaptive to the inspected signal.

For each selected time frame, by taking the normalized dot product between the exact wavelet coefficients and the inspected signal, a fitness index can be obtained to evaluate the degree of matching. The index is calculated using a cosine function of two vectors:

$$\cos(\vec{C}, \vec{X}) = \frac{\sum_{i=1}^{n} c_i x_i}{\sqrt{\sum_{i=1}^{n} c_i^2}\sqrt{\sum_{i=1}^{n} x_i^2}} \quad (1 \leqslant n \leqslant N), \tag{8}$$

where $\vec{C}$ and $\vec{X}$ stand for the vectors of the wavelet coefficients and the portion of the inspected signal that is required to be matched respectively. The variables c_i and x_i represent the elements of the vectors. N is the number of data. The calculated index from the fitness function provides a measure to evaluate the similarity of the two vectors not only in their magnitudes but also in their geometrical shapes. The higher the index of the fitness function, the more similar are the derived wavelet and the portion of the inspected signal. The index of the cosine function approaches to 1 indicates a prefect match, whilst the index approaches to zero shows a mismatch.

3.4. Optimization of exact wavelets in simulated signals

Based on the aforementioned theories, a series of exact wavelets were derived and applied to a simulated Doppler signal for testing their effectiveness. The simulated signal is shown in the bottom diagram of Fig. 5. The diagram shows that the frequency of the simulated signal increases with time, and a temporal impulsive feature is presented at the first portion of the simulated signal. To derive the exact wavelets for matching the simulated Doppler signal, the crossover and mutation operators commonly used in genetic algorithms were employed for the optimization process [25]. The control parameters used in the optimization process are listed in Table 1. To ease the computational requirements and to increase the efficiency of the process, the population scale is set to 100, the number of iteration to 400, the probability of crossover to 0.6 and the probability of mutation to 0.02. The variables ς, Ω and θ, and the scale a are coded and decoded using the binary coding/decoding mechanism. The relationship of the precision in calculation and the coded variables is defined according to Goldberg's suggestion [25] as

$$\delta = \frac{U_{max} - U_{min}}{2^l - 1}, \tag{9}$$

where δ stands for the required precision, U_{max} and U_{min} represent the maximum and the minimum values of the variable that is being coded or decoded respectively, and l refers to the code length. Once the required precision and the range of each variable are known, the code length can be determined using Eq. (9).

By using the control parameters as listed in Table 1, the optimized exact wavelets that correspond to every point or unit of time of the signal have been derived. The results generated by the exact wavelets are compared to the simulated signal and displayed in Fig. 5. The results clearly show that both the transient features and the shape of the resulting coefficients (top diagram) closely match with the simulated signal (bottom diagram). The middle diagram of Fig. 5 displays the corresponding scales which reflect the frequency of the signal at each unit of time. This scale distribution diagram reveals the frequency variation of the simulated signal precisely. For the

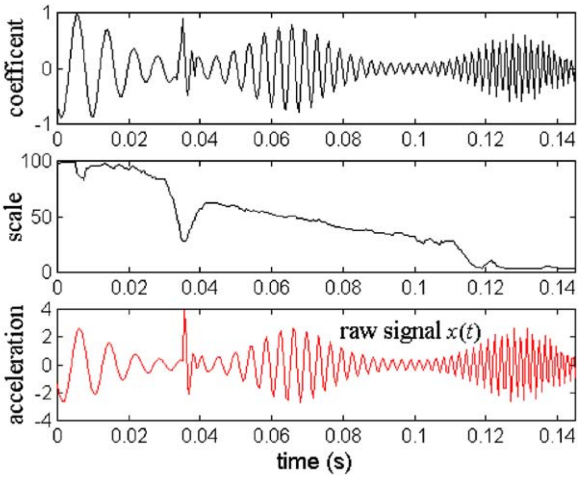

Fig. 5. The transient acceleration and scale information of the simulated signal.

Table 1
The control parameters that are used in the genetic algorithm for optimizing and deriving the exact wavelets

Number order	Items	Values of parameters
1	Population scale	100
2	Terminal iteration times	400
3	Probability of crossover	0.6
4	Probability of mutation	0.02
5	Calculation precision	
	$\varsigma \in [1, 100]$	$3e^{-3}$
	$\Omega \in [1, 256]$	1
	$\theta \in [0, 2\pi]$	$1e^{-3}$
	$a \in [10, 200]$	1
6	Length of binary code	
	ς	15
	Ω	8
	θ	13
	a	8

purpose of comparison, the 3D distribution maps of the time-scale for the simulated signal that are generated by exact wavelet analysis and conventional Morlet CWT are displayed in Figs. 6(a) and (b) respectively. Note that exact wavelet analysis shows a clear and precise distribution that well matches the simulated signal. In contrast, the conventional Morlet CWT generates a smear distribution map that is full of overlapping and distortions. The time and frequency properties of the simulated signal can hardly be identified from the map generated by Morlet CWT as shown in Fig. 6(b). The machine operators could not draw any useful decision on machine fault diagnosis from such inconclusive result.

It is worth noting that the result generated by exact wavelet analysis is totally different from the results that are generated by wavelet ridge extraction or other adaptive wavelets. In the result of exact wavelet analysis, there is only one transient amplitude of coefficient and one transient scale at each unit of time. This one to one corresponding relationship between the generated coefficients and the inspected signal is a powerful feature and uniquely offered by exact wavelet analysis. Hence, exact wavelet analysis is useful in analyzing the non-linear and non-stationary temporal signals that are caused by randomly occurring faults.

4. Applications of exact wavelet analysis in industrial machine fault diagnosis

To investigate the effectiveness of exact wavelet analysis in industrial machine fault diagnosis, a series of vibration signals collected from a real machine were analyzed for detecting possible faults occurring during the operation of the machine. The real machine is a motor-pump drive system as shown in Fig. 7. It is composed of a variable speed AC motor, a hydraulic pump, flexible couplers, a number of ball bearings and journal bearings, a gear coupler, shafts, and a flywheel for balancing. The load of the pump can be adjusted via a variable displacement valve. In the

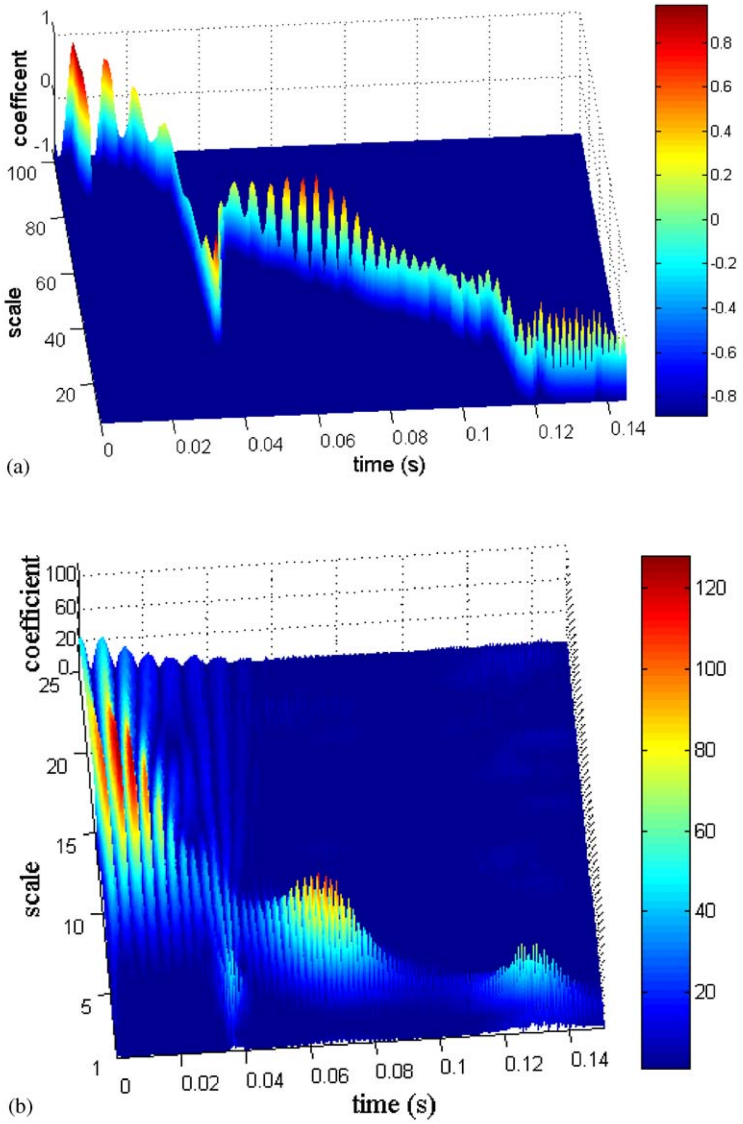

Fig. 6. The simulated signal's time-scale distribution maps as generated by exact wavelet analysis and the conventional Morlet CWT. (a) Using the exact wavelet analysis. (b) Using the conventional Morlet CWT.

experiments, two different kinds of faults had occurred in ball bearing 1 which is labelled in Fig. 7. To collect the vibration signals, accelerometers were installed in the bearing house in both the radial and axial directions of the bearing housing. The shaft of the bearing ran at 23.3 Hz or 1398 r.p.m. Usually, faults that occur in ball or roller bearings are related to their defective inner-races, outer-races, rolling elements or cages [26]. In this study, defects occurred in the inner-race and outer-race of the ball bearing 1 at different times. The geometric parameters of the ball bearing 1 are listed in Table 2. Based on the geometric parameters and the rotational speed of ball

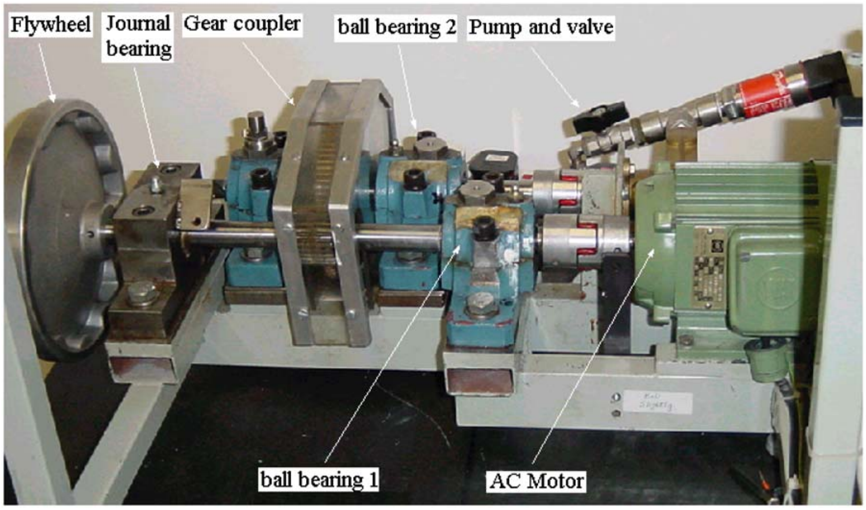

Fig. 7. The configurations of the motor-pump rotary machine.

Table 2
The geometric parameters and the characteristic frequencies of ball bearing 1

Ball diameter	7.5 mm
Pitch diameter	39.4 mm
Contact angle	0°
Number of rolling elements	13
Ball-passing frequency inner-race (BPFI)	179 Hz (5.6 ms)
Ball-passing frequency outer-race (BPFO)	122 Hz (8.2 ms)

bearing 1, its characteristic frequencies can be calculated [5]. Its ball-passing frequency inner-race (BPFI) and the ball-passing frequency outer-race (BPFO) are estimated as 179 and 122 Hz, respectively.

The existence of faults often causes impulses or impacts that are resident in the vibration signal [5]. When the defect is minor, the vibration energy that is generated by the defect is small. Hence, the fault-related impulses are difficult to distinguish from the broad spectrum, because they may be overwhelmed or even buried by other larger structural vibrations and background noise [27]. Exact wavelet analysis was applied to detect and extract such kind of buried fault-related impulses from the vibration signal collected at the bearing's housing.

When the bearing was operating normally, the vibration signal was collected. The temporal signal of vibration is shown in the bottom diagram of Fig. 8, The results of exact wavelet analysis, including the transient amplitude of the coefficient and the scale values for each unit of time, are plotted in the top and middle diagrams of Fig. 8 respectively. Note that no distinctive high magnitude coefficients, which are related to impulses caused by faults, can be found from the top diagram. Moreover, the dominant scale of the analyzed results (middle diagram) is around 160–180. Therefore, most of the frequencies that are embedded in the inspected signal are relatively low in frequency level.

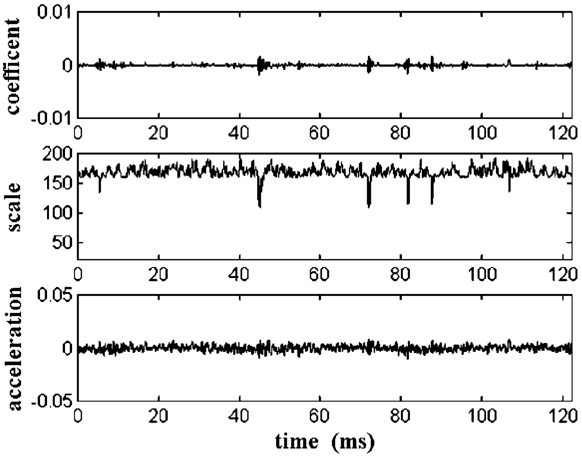

Fig. 8. Vibration signals and the results of the bearing when running normally.

Later, an inner-race defect was discovered in the bearing. The raw vibration signals were collected when the defect was in moderate and then serious conditions as shown in the bottom diagrams of Figs. 9(a) and (b) respectively. Finally, the bearing required an overhaul. Again, exact wavelet analysis was used to analyze the vibration signals. When the inner-race defect was in moderate condition, the results, which include the transient magnitude of the coefficient and the scale values for each unit of time, are plotted in the top and middle diagrams of Figs. 9(a) respectively. Similarly, when the defect was in serious condition, the transient magnitude of the coefficient and the scale values for each unit of time, are plotted in the top and middle diagrams of Fig. 9(b).

Note that in the top diagrams of Figs. 9(a) and (b), high magnitude coefficients, which are related to faulty impulses, can be clearly revealed using exact wavelet analysis even though noise is embedded in the inspected signals. In the top diagrams, quasi-periodic intervals that are approximately equal to 5.6 ms can be found in the transient magnitudes of the coefficients. Similar intervals can also be identified in the middle diagrams of Figs. 9(a) and (b) for scale distribution, particularly at the scales with low values. Such quasi-periodic intervals are equivalent to the inverse of the ball-passing frequency inner-race (BPFI) which is 179 Hz as listed in Table 2. Hence, it can be concluded that the impulses are caused by the inner-race defect. It is worth to note here that the fault-related impulses are quasi-periodic. The authors have done many tests and consultancies in machine fault diagnosis for industries. Obvious periodic impacts are hardly seen unless the damage of the defective component is very serious and near fatal breakdown. The idea of machine fault diagnosis is to detect the occurrence of faults as soon as possible or in other words, when the damage caused by faults is not serious. Therefore, the detection of quasi-periodic impulses is important for obtaining advance warning prior to catastrophe. Exact wavelet analysis does help to obtain such advance warning.

Another interesting observation is that the fault-related impulses always posses concentrated vibration energy in a short burst. That is, most of the fault-related impulses have high frequency content or occur at low scale values. Compared to the normal condition, as shown in the middle diagram of Fig. 8, the scale levels have dropped from an average of 160 to as low as 30, depending

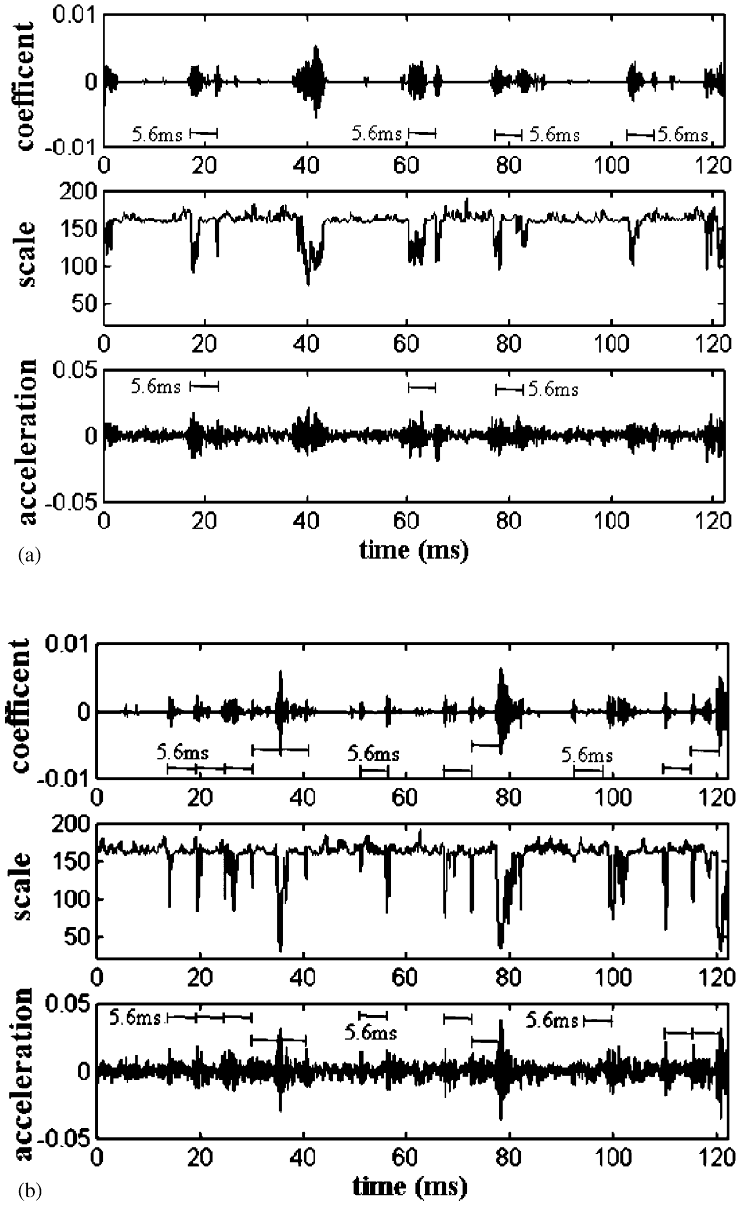

Fig. 9. Vibration signals and the results of the bearing with moderate and serious inner-race defects. (a) Moderate inner-race defect condition. (b) Serious inner-race defect condition.

on the severity of the defect. To investigate the time–frequency properties of the vibrations of a normal and a defective bearing, the distributions of the number of data points against the scale values smaller than 160 are plotted in Fig. 10. Under normal running condition, nearly all of the data are located at scale values above 160. Only a few data points exist between the scale values from 160 to 110. Under anomalous running conditions, the scale values have dropped

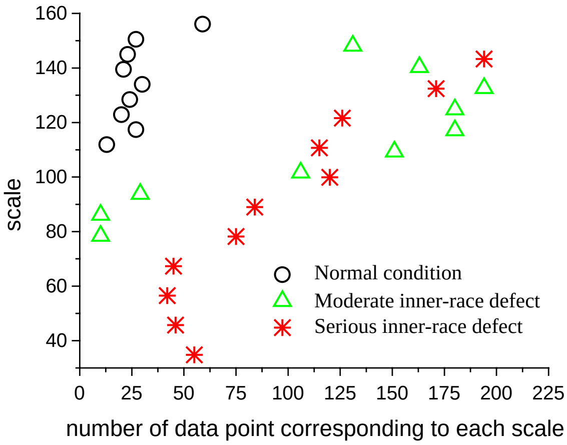

Fig. 10. Distributions of scale levels for normal and inner-race defect conditions.

significantly. The statistical analysis shows that the total number of data points that are located below the scale value of 160 is 158, 1144 and 1476 for normal, moderate inner-race defect, and serious inner-race defect running conditions respectively. The statistical results disclose that the more serious the damage, the more fault-related impulses will occur at significantly high frequency zone or low scale region. This phenomenon can be explained by the fact that when a defective surface comes into contact with another surface, a high intensity of vibration energy is emitted for a short time or at a high frequency. Such phenomenon provides the machine operator with an extra piece of evidence to judge whether a fault has occurred in a running ball bearing.

To ensure that exact wavelet analysis is also effective for other type of bearing faults, an artificial outer-race defect was introduced to the bearing after its overhaul by replacing a new inner-race. Vibration signals were collected for both moderate and serious outer-race defect conditions, as shown in the bottom diagrams of Figs. 11(a) and (b) respectively. Both figures also show the transient scales and the transient magnitude of the coefficients in the middle and top diagrams. As in the case of the inner-race defect, quasi-periodic intervals equal to 8.2 ms can be found in the top and middle diagrams of both figures. These quasi-periodic intervals are equivalent to the inverse of the ball-passing frequency outer-race (BPFO) which is 122 Hz as listed in Table 2. Hence, it can be concluded that the impulses are caused by the outer-race defect. Moreover, nearly all the impulses occurred at much lower scale values (the middle diagrams of Figs. 11(a) and (b)) than the scale values of the normal condition (the middle diagram of Fig. 8). The distributions of the number of data points against the scale values smaller than 160 are plotted in Fig. 12. For the moderate and serious outer-race defect conditions, the impulses are located mostly below the scale value 160 to as low as 70. Whilst for the normal condition, the lowest vibration is at the scale 110. According to the statistical analysis, the total number of vibration data that are located below the scale value 160 is 627 and 727 for moderate and serious outer-race defect conditions respectively. Obviously, the statistical results again reveal that the more serious the defect, the more number of impulses appear at high frequency zone or low scale region. The operator can use the scale distribution as a simple and instant method for evaluating

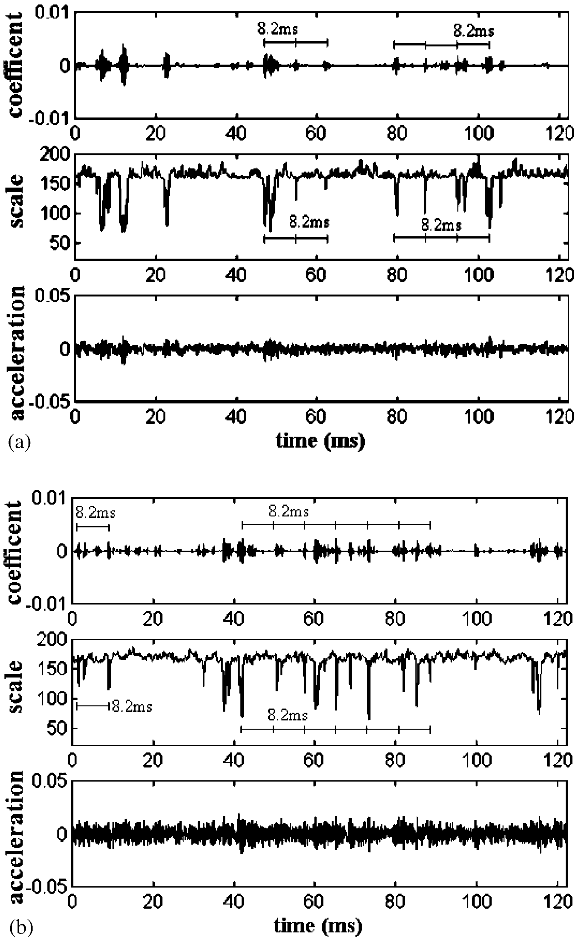

Fig. 11. Vibration signals and the results of the bearing with moderate and serious outer-race defects. (a) Moderate outer-race defect condition. (b) Serious outer-race defect condition.

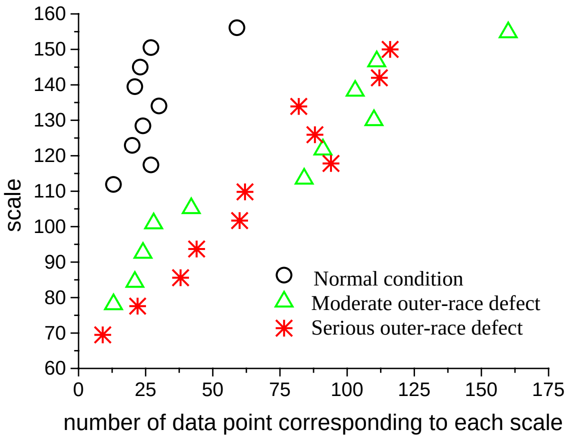

Fig. 12. Distributions of scale levels for normal and outer-race defect conditions.

the running condition of any ball bearing. Conventional CWTs cannot provide such an effective and prompt method.

From the results of the simulated test and the experiments performed on the bearing of the motor-pump drive during the normal, the inner-race defect and the outer-race defect conditions, exact wavelet analysis is found to be effective in fault diagnosis. It helps machine operators not only in detecting the existence of faults by using the scale distribution plot, but also in identifying the causes of faults by using the information of the time intervals which is provided by both the transient coefficient and the scale diagrams. The statistical analysis of the distribution of the total number of data points versus the scales provides additional evidence for evaluating the severity of the damage that has been caused by the faults to the machine.

5. Conclusions

An innovative wavelet called *exact wavelet analysis* has been designed to enhance the robustness of vibration-based machine fault diagnosis. The deficiencies of conventional CWTs in machine fault diagnosis have been thoroughly investigated. The newly derived exact wavelets are optimized so that they can precisely reveal the time and frequency properties of the inspected signal. The effectiveness of exact wavelet analysis has been demonstrated by both simulated and practical experiments. From the derived theories and the experimental results, concluding remarks are drawn as follows.

The effects caused by overlapping and distortion have been significantly reduced (as shown in Figs. 5 and 6(a)) by exact wavelet analysis. These benefits cannot be achieved using conventional CWTs (Fig. 6(b)). Exact wavelet analysis can provide a precise and clear transient plot of the coefficients and scales. The results that are generated by exact wavelet analysis can closely match the time and frequency properties of the inspected signal. The results can *exactly* distribute at their expected time and scale locations as originally appeared in the inspected signal.

By varying the variables ς, Ω, and θ, multiple mother wavelets can be derived. The existence of multiple mother wavelets will generate various kinds of daughter wavelets that are adaptive to the characteristics of the inspected signal. Hence, exact wavelet analysis could be superior to other optimized wavelets, which use only one single mother wavelet to characterize the inspected signal. The distortion and dilution to the inspected signal that commonly occur in other wavelets have been significantly reduced.

Both the optimization of the wavelet coefficients and the scales, and the satisfaction of the admissibility conditions, have been fully considered and satisfied during the design of exact wavelet analysis. Such a thorough consideration ensures that the derived exact wavelets can match the transient properties of the inspected signal as closely as possible. Thus, exact wavelet analysis possesses a powerful ability in the extraction of both stationary and transitory features.

The fault-related impulses that are caused by defects can be revealed in both the transient coefficient and scale plots that are generated by exact wavelet analysis. By using these plots, machine operators can easily detect the existence of faults and identify the causes of the faults. Such convenience is not available from conventional CWTs.

The distribution of scales derived under different machine running conditions can also provide an important clue in the evaluation of different running conditions of a machine. Normal and anomalous running conditions reveal significantly different scale distributions. A machine that has a fault has much lower scale values. The severe the fault, the higher the vibration energy contained in the fault-related impulses, and the more the impulses appear at high frequency zone or low scale region.

Both the theoretical analyses and the experimental results show that exact wavelet analysis is an effective tool for vibration-based machine fault diagnosis. Exact wavelet analysis can be used to extract fault-related features that exhibit both stationary and non-stationary characteristics, making it particularly suitable for detecting randomly occurring faults. Exact wavelet analysis provides an unambiguous diagnostic ability to machine operators for detecting the existence of faults and determining the severity of deterioration of a defective machine.

Acknowledgements

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 1046/00E).

References

[1] W.J. Staszewski, Structural and mechanical damage detection using wavelets, *The Shock and Vibration Digest* 30 (1998) 457–472.

[2] B.A. Paya, I.I. Esat, M.N. Badi, Artificial neural network based fault diagnostics of rotating machinery using wavelet transforms as a pre-processor, *Mechanical System and Signal Processing* 11 (5) (1997) 751–759.

[3] A.C. Okafor, A. Dutta, Structural damage detection in beams by wavelet transforms, *Smart Materials and Structures* 9 (6) (2000) 906–917.

[4] W.J. Wang, Wavelet for detecting mechanical faults with high sensitivity, *Mechanical System and Signal Processing* 15 (4) (2001) 685–696.

[5] P. Tse, Y.H. Peng, R. Yam, Wavelet analysis and envelope detection for rolling element bearing fault diagnosis—their effectiveness and flexibility, *Transactions of the American Society of Mechanical Engineers, Journal of Vibration and Acoustics* 123 (3) (2001) 303–310.

[6] C. Cempel, *Vibroacoustics Condition Monitoring, Ellis Horwood Series in Mechanical Engineering*, Ellis Horwood Ltd., Chichester, 1991.

[7] R.B. Randall, Frequency Analysis, Brüel & Kjaer Ltd., Nærum, pp. 263–269.

[8] B. Liu, S.F. Ling, On the selection of informative wavelets for machinery diagnosis, *Mechanical Systems and Signal Processing* 13 (1) (1999) 145–162.

[9] D.E. Newland, Ridge and phase identification in the frequency analysis of transient signals by harmonic wavelets, *Transactions of the American Society of Mechanical Engineers, Journal of Vibration and Acoustics* 121 (1999) 149–155.

[10] S. Mallat, Z. Zhang, Matching pursuit with time-frequency dictionaries, *IEEE Transactions on Signal Processing* 41 (12) (1993) 3397–3415.

[11] B.A. Telfer, H.H. Szu, G.J. Dobeck, J.P. Garcia, H. Ko, A. Dubey, N. Witherspoon, Adaptive wavelet classification of acoustic and backscatter and imagery, *Optical Engineering* 33 (1994) 2192–2203.

[12] H.H. Szu, B.A. Telfer, S. Kadambe, Neural network adaptive wavelets for signal representation and classification, *Optical Engineering* 31 (1992) 1907–1916.

[13] C.M. Kocur, S.K. Rogers, L.R. Myers, T. Burns, M. Kabrisky, et al., Using neural networks to select wavelet features for breast cancer diagnosis, *IEEE Engineering in Medicine and Biology* 95–102 (May/June) (1996) 108.

[14] G. Tagliarini, E. Page, Genetic algorithms for adaptive wavelet design, SPIE Wavelet Applications III 2762, 8–12 April, Orlando, FL, 1996, pp. 82–93.

[15] R.K.H. Galvao, T. Yoneyama, T.N. Rabello, Signal representation by adaptive biased wavelet expansions, *Digital Signal Processing* 9 (1999) 225–240.

[16] A.R.F.D. Silva, Evolutionary based methods for adaptive signal representation, *Signal Processing* 81 (5) (2001) 927–944.

[17] I. Daubechies, *Ten Lectures on Wavelets*, Science for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992, pp. 1–351.

[18] P. Tse, W.X. Yang, A new wavelet transform for eliminating problems usually occurring in conventional wavelet transforms used for fault diagnosis, *The Ninth International Congress on Sound and Vibration, ICSV'9*, Orlando, Florida, Paper # 465, CD-ROM version, 2002.

[19] P. Tse, W.X. Yang, The practical use of wavelet transforms and their limitations in machine fault diagnosis, *International Symposium on Machine Condition Monitoring and Diagnosis invited paper*, Tokyo, Japan, 2002, pp. 9–16.

[20] B.L. Borde, New wavelet class for fine structure identification, *SPIE Wavelet Applications II* 2491–2499 (1995) 1073–1085.

[21] B.L. Borde, Generic explicit wavelet tap derivation, *SPIE Wavelet Applications III* 2762 (1996) 94–104.

[22] D.G. Luenberger, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[23] R. Fletcher, *Practical Methods of Optimisation*, Wiley, New York, 1986.

[24] A.A. Rabow, H.A. Scheraga, Lattice neural network minimization application of neural network of optimisation for locating the global-minimization conformations of proteins, *Journal of Molecular Biology* 232 (4) (1993) 1157–1168.

[25] D.E. Goldberg, *Genetic Algorithms in Search, Optimisation, and Machine Learning*, Addison Wesley, Reading, MA, 1989.

[26] D. Bently, Predictive maintenance through the monitoring and diagnostics of rolling element bearings, Applications Note, ANO44, Bently Nevada Co., 1989, pp. 2–8.

[27] D. Ho, R.B. Randall, Optimisation of bearing diagnostic techniques using simulated and actual bearing fault signals, *Mechanical System and Signal Processing* 14 (5) (2000) 763–788.