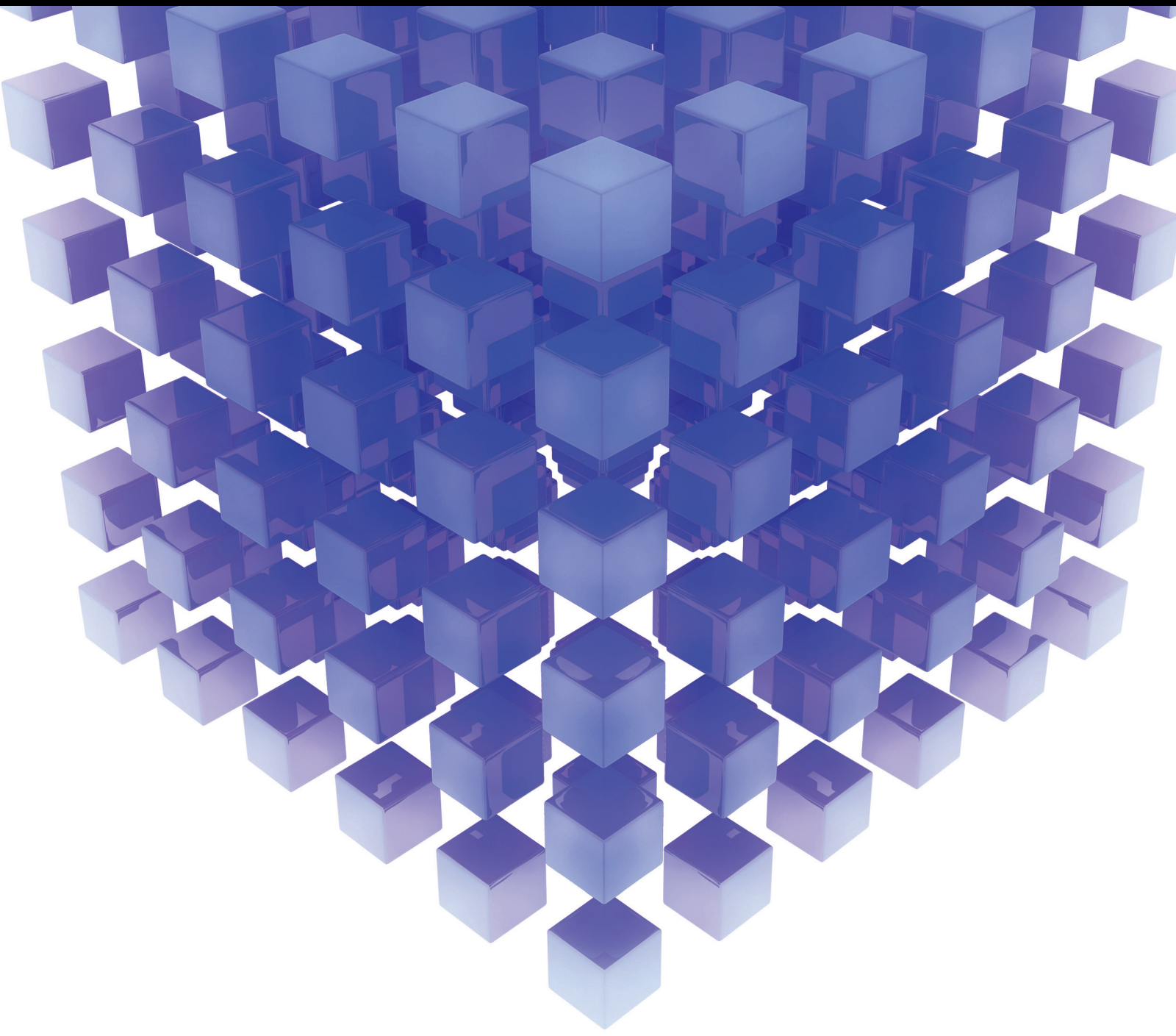
Mathematical Problems in Engineering

Mathematical Methods and Modeling in Machine Fault Diagnosis

Guest Editors: Ruqiang Yan, Xuefeng Chen, Weihua Li, and Shuangwen Sheng



Mathematical Methods and Modeling in Machine Fault Diagnosis

Mathematical Problems in Engineering

Mathematical Methods and Modeling in Machine Fault Diagnosis

Guest Editors: Ruqiang Yan, Xuefeng Chen, Weihua Li, and Shuangwen Sheng

Editorial Board

Contents

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 516590, 3 pages
http://dx.doi.org/10.1155/2014/516590

Editorial

Mathematical Methods and Modeling in Machine Fault Diagnosis

Ruqiang Yan,[1] Xuefeng Chen,[2] Weihua Li,[3] and Shuangwen Sheng[4]

[1] *School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China*
[2] *School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China*
[3] *School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510640, China*
[4] *National Renewable Energy Laboratory, Golden, CO 80401, USA*

Correspondence should be addressed to Ruqiang Yan; ruqiang@seu.edu.cn

Received 20 August 2014; Accepted 20 August 2014; Published 18 December 2014

Modern mathematics has often been utilized as an effective tool to model mechanical equipment so that their dynamic characteristics can be studied analytically. This will help identify potential failures of mechanical equipment by observing change in the equipment's dynamic parameters. On the other hand, dynamic signals are also important and provide reliable information about the equipment's working status. Modern mathematics has also provided us with a systematic way to design and implement various signal processing methods, which are used to analyze these dynamic signals, and to enhance intrinsic signal components that are directly related to machine failures. This special issue is aimed at stimulating not only new insights on mathematical methods for modeling but also recently developed signal processing methods, such as sparse decomposition with potential applications in machine fault diagnosis. The papers included in this special issue provide a glimpse into some of the research and applications in the field of machine fault diagnosis through applications of the modern mathematical methods.

Manufacturing systems are often operated under high pressure, at high temperatures, with fast-flowing material and complex mechanisms. Progressive faults can be caused by deterioration of the operating environment or aging and show stochastic properties. X. Gong et al. explored an adaptive maintenance model to detect progressive faults. The model is used to monitor the environmental states of the systems and compares the inputs/outputs and presettings to detect faults. Experiments performed on the process environment of an air separation plant verified the effectiveness of the maintenance model.

Rotating machines and/or machine components have been widely used in modern factories. Timely monitoring and diagnosing their working conditions is critical to avoid possible losses of production due to machine breakdown, as these machines are always running under varying speeds and loading conditions, which make the vibration signal more complicated. To address this issue, S. Zhang and W. Li explored an unsupervised neighborhood preserving embedding (NPE) algorithm with application to bearing fault diagnosis. Their study indicates the NPE is capable of extracting the most discriminative features for classification. Furthermore, when combined with a self-organizing map, the bearing degradation process can be assessed accurately. F. Wang et al. proposed a new tensor manifold method to realize the bearing fault feature extraction. The time-frequency characteristics of the signals are extracted using the tensor manifold. The proposed method can reduce the information redundancy and information loss and effectively distinguish different bearing fault states. Q. He et al. acquired sensitive features through a combination of local discriminant bases (LDB) and locality preserving projections. The proposed feature extraction method combines the merits of these two techniques and extracts the inherent pattern structure embedded in the discriminatory features. The new feature

not only considers the static discriminatory wavelet packet node features themselves, but also considers the dynamic sensitive class pattern structure embedded in the samples. The proposed feature displays valuable benefits for data classification, and its effectiveness is verified by case studies on vibration data-based classification of bearing fault types and severities. W. Cheng et al. reported a feature selection criterion, mean impact variance (MIVAR), which can be used to determine which feature is more suitable for the artificial neural network-based bearing fault classification. The MIVAR values of all the features are calculated by changing the input vectors and then measuring the differences of the output vectors after the training process of the back propagation neural network. It is proved that using the features with higher MIVAR values can lead to a higher recognition rate and the corresponding performance is as good as that of a traditional feature selection algorithm, such as the principal component analysis (PCA) algorithm. In another study, H. Wang et al. developed a new kind of peak-based strategy to enhance the weak bearing fault detection. A peak-based piecewise recombination is proposed to convert middle frequency components into low frequency ones; then the vibration signal becomes so smooth that its sparseness in the wavelet domain will improve significantly. This helps eliminate interference noise and detect weak bearing faults. H. Zhang et al. proposed a new stochastic resonance (SR) model for bearing fault diagnosis. The new SR model has a joint Woods-Saxon and Gaussian potential whose parameters are not coupled and thus easily tuned to optimize the output signal-to-noise ratio (SNR). In addition, a smoother potential bottom and a steeper potential wall lead to stable particle motion within each potential well and thus avoid the unexpected noise. The novel bistable SR model is verified to be capable of offering a higher output SNR and a wider bandwidth in weak signal detection. S. Dong et al. presented a hybrid method to solve the rotating machinery fault diagnosis problem, which is based on PCA to extract the characteristic features and the Morlet kernel support vector machine to achieve the fault classification. The proposed method makes good use of the advantage of all parts together to obtain better recognition accuracy. H. Sun et al. studied a fault diagnosis method for rotating machinery based on a multiwavelet-adaptive threshold denoising and mutation particle swarm optimization algorithm (MPSO). A Geronimo, Hardin, and Massopust (GHM) multiwavelet was employed for extracting weak fault features under background noise, and the method of adaptively selecting the appropriate threshold for the multiwavelet with an energy ratio of a multiwavelet coefficient was presented. An MPSO algorithm with an adaptive inertia weight adjustment and a particle mutation was proposed for condition identification. Practical examples of fault diagnosis for rolling element bearings verified the effectiveness of the proposed method. W. Fan et al. proposed a novel method for machinery fault diagnosis by combining the wavelet basis and majorization-minimization algorithm. With the proposed method, transients hidden in the noisy signal can be converted into sparse coefficients; thus the transients can be detected sparsely. The effectiveness of the proposed method is verified by both simulated and measured gearbox vibration signals. Results show that the proposed method outperforms the method based on split-augmented Lagrangian shrinkage algorithm in convergence and detection effect.

The aeroengine is one of the key components in an aircraft and its reliability directly affects flight safety. Rub-impact caused by decreased clearance between the rotor and stator in an aeroengine generates unexpected vibrations, making the aeroengine not function well and even causing catastrophic consequences. Identifying the rub-impact fault at its early stage in an aeroengine becomes a critical task. J. Cui et al. developed an integrated approach, based on the improved LDB and support vector machine, for aeroengine fault diagnosis. The experimental results verified that the developed approach is able to classify different aeroengine working conditions. Predicting fatigue crack propagation of aircraft components under service loading is necessary to ensure flight safety. S. Jiang et al. developed a theoretical model to predict fatigue crack growth behavior under the single overload, in which crack closure and a plastic zone concept are considered. The model was validated in D16 aluminum alloy and 350WT steel subjected to several different loading spectra, and the predictions matched experimental data well.

The blade is a key component of the centrifugal compressor. However, crack and fatigue failure can often occur as the blade is subjected to centrifugal forces, gas pressure, friction force, and so on. Thus, it is important to have an early warning for blade cracking. H. Li et al. used pressure pulsation information to diagnose blade crack from a blade vibration transfer process analysis. A dynamic strain sensor is installed on the blade to determine the crack characteristic frequency. The results show that this method can be helpful for blade crack classification in centrifugal compressors.

Rail track inspection and maintenance are key factors in keeping trains operating safely. J. Zhao et al. presented a two-dimensional impact reconstruction method to perform online inspection of rails and to find defects. The method utilizes preprocessing technology to convert time domain vertical vibration signals, acquired by a wireless sensor network, to space signals. The modern time frequency analysis method is improved to reconstruct the obtained multisensor information. Then, image fusion processing technology, based on spectrum threshold processing and node color labeling, is introduced to reduce the noise, blank the periodic impact signals caused by rail joints and locomotive running gear, and convert the aperiodic impact signals caused by rail defects to partial periodic impact signals. This method can be used to do an online analysis of the vertical vibration signals of a train when it is running, extract the aperiodic impact features caused by the rail defects, perform the online inspection, and locate rail defects.

Fiber optic current transducers (FOCTs) are the basic components of power systems. To ensure low failure and high reliability of FOCTs, it is vital to study methods of condition monitoring and fault diagnosis in such transducers. As a complement to the frequency domain analysis, L. Wang et al. analyzed time domain and frequency domain features of fiber optic current transformers' measurement data, established correspondence between the physical characteristics of key components in transformer and data features, and then built

a diagnostic analysis model based on Allan variance. The fiber optic current transformer's health state can be determined from the Allan variance calculation results.

Magnetic flux leakage (MFL) is an efficient method for detecting pipeline flaws. Z. Wu et al. proposed a new adaptive channel equalization approach for processing a MFL signal prior to flaw characterization. The approach performs channel equalization by using single layer neural networks, and the fast-learning algorithm of an extreme learning machine is used to achieve excellent processing speed. Focusing on the signal of a flaw, a PCA-based flaw detecting algorithm is given to locate the flaw signal, which is verified from both theoretical and simulation studies.

Polymer matrix composite materials have been developed rapidly to meet the demands for better materials with higher standards of performance and reliability in structures and machines. X. Zhu et al. employed a viscoplastic constitutive model in the micromechanical method, based on a generalized model of cells, to analyze the inelastic, rate dependent stress-strain response of fiber-reinforced polymer matrix composites with three different microstructures at different fiber off-axis angle conditions. Acceptable agreement is observed between the model predictions and experimental results found in the literature. The results show that the stress-strain curves are sensitive to the strain rate and the microstructure parameters play an important role in the behavior of a polymer matrix.

Acknowledgments

As our editorial work comes to an end, we would like to express our deep appreciation to all the authors who supported this special issue by contributing papers. We are also grateful to all the reviewers for their insightful and constructive comments.

Ruqiang Yan
Xuefeng Chen
Weihua Li
Shuangwen Sheng

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 862065, 15 pages
http://dx.doi.org/10.1155/2014/862065

Research Article

Pressure Pulsation Signal Analysis for Centrifugal Compressor Blade Crack Determination

Hongkun Li,[1] Xuefeng Zhang,[1] Xiaowen Zhang,[1] Shuhua Yang,[2] and Fujian Xu[1]

[1] *School of Mechanical Engineering, Dalian University of Technology, No. 2 Linggong Road, Dalian 116024, China*
[2] *Shenyang Blower Works Group Corporation, Shenyang 110869, China*

Correspondence should be addressed to Hongkun Li; lihk@dlut.edu.cn

Received 28 March 2014; Revised 31 May 2014; Accepted 16 June 2014; Published 19 August 2014

Academic Editor: Ruqiang Yan

Blade is a key piece of component for centrifugal compressor. But blade crack could usually occur as blade suffers from the effect of centrifugal forces, gas pressure, friction force, and so on. It could lead to blade failure and centrifugal compressor closing down. Therefore, it is important for blade crack early warning. It is difficult to determine blade crack as the information is weak. In this research, a pressure pulsation (PP) sensor installed in vicinity to the crack area is used to determine blade crack according to blade vibration transfer process analysis. As it cannot show the blade crack information clearly, signal analysis and empirical mode decomposition (EMD) are investigated for feature extraction and early warning. Firstly, signal filter is carried on PP signal around blade passing frequency (BPF) based on working process analysis. Then, envelope analysis is carried on to filter the BPF. In the end, EMD is carried on to determine the characteristic frequency (CF) for blade crack. Dynamic strain sensor is installed on the blade to determine the crack CF. Simulation and experimental investigation are carried on to verify the effectiveness of this method. The results show that this method can be helpful for blade crack classification for centrifugal compressors.

1. Introduction

With the development of the society, centrifugal compressor has been widely used in modern industry such as petroleum, chemicals, metallurgy, and aerospace field as an important fluid machine [1]. Meanwhile, centrifugal compressors are developing to be large in scale, high in speed, and automatic in operation [2]. However, the blade failure usually emerges. As the most important part, the impeller transforms kinetic energy into pressure energy. But the impeller suffers from the effect of centrifugal forces, gas pressure, and friction force which usually lead to cracks. According to statistical analysis, 65% centrifugal compressor malfunctions are closely related to the blades. In addition, 40% blade fatigue failures are not fully understood so far [3]. Examples of blade cracks are shown in Figure 1. Fluid-induced vibration is an important factor for blade fatigue failures. It contains acoustic resonance, unsteady flow, rotating stalls, and flutter [4, 5]. Due to the high-velocity flow through the centrifugal compressor and rotating impeller, high-pressure fluctuations occur in the cavity of compressor which could lead the impeller to irregular vibration. Pressure fluctuation acts on the impeller, leading to stress convergence and cracks in the blades. The growing crack will cause blade failure, which results in catastrophes.

There are many reasons for cracks on the blades of compressor. Blade cracks are mainly associated with the detection of material, production process, working condition, and high cycle fatigue. So far, researches were mainly concentrated on the defects in the material, processing, and manufacture of impeller causing high fatigue failure. Lourenço investigated the failure of blades [6]. Kermanpur et al. analyzed the failure mechanism of compressor blades made of Ti6Al4V. The results showed fretting fatigue mechanism is the main cause of several premature failures of Ti6Al4V alloyed compressor blades [7]. In recent years, blade cracks caused by excessive alternating stress induced by air-excited vibration have drawn more and more attention from the researchers.

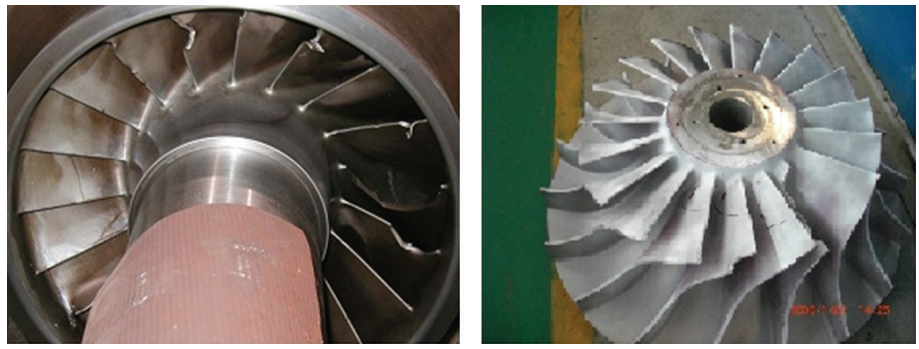


Figure 1: Pictures of centrifugal compressor blade cracks.

In 2007, Eisinger studied the acoustic fatigue which is the coincidence of impeller structural and cavity acoustic modes. The results indicated that the acoustic fatigue would significantly increase the amplitude of vibration and damage the blades [8]. Investigation on alternating stress can be helpful to prevent or reduce the damage from blade cracks [9]. Therefore, condition monitoring and pattern classification are important to prevent blades from failure as well as blade crack detection which could ensure safe operation of the compressor.

It is well known that blade cracks will result in breakdown or even serious accidents of the whole set for centrifugal compressor. It can even lead to heavy losses for a factory. Moreover, personal safety must be considered because the tangential velocity of breakdown blade can be up to 450 m/s. Therefore, incipient classification of blade crack becomes more and more important than ever before. Traditionally, displacement sensors are introduced to monitor shaft vibration. Meanwhile, vibration-based condition monitoring is also used in shaft crack classification [10, 11]. But it is difficult to recognize shaft cracks only by vibration signals. Moreover, it is impossible to provide any information to characterize blade crack condition from the shaft vibration signals, making blade crack classification more difficult than shaft crack identification. Different methods for blade condition classification have been investigated by many researchers. Liu et al. studied the malfunction identification method of fan blade crack classification by using wavelet packet analysis [12]. Though the structure is similar to centrifugal compressor and fan, centrifugal compressor has good stiffness as the typical difference. Rama Rao and Dutta studied blade crack condition classification for gas turbine blade recognition by using vibration signal information [13]. Yang et al. proposed the auditory spectrum feature extraction using the support vector machine to identify the malfunction of fans [14]. Witek studied the experimental crack propagation for gas turbine blades via vibration signals in laboratory but it was not in a close-loop test-rig [15]. At the same time, some researchers studied wind turbine blade crack classification by using wavelet analysis, scalogram, and so on [16–18]. But it is different from centrifugal compressor blade working condition in speed and load. All these investigations are helpful for blade crack classification, but further study for early warning of centrifugal compressor blade is required. At the same time, air flow experiment is more important for blade condition analysis in real working conditions.

Pressure pulsation (PP) generated by the interference between rotating blades and the stationary vanes contains much information about the blade working conditions and has been used for blade status conditions analysis [19]. However, the crack information in PP signal is weak, so it is difficult to identify patterns just according to time or frequency information, especially for the incipient blade crack condition. Further feature extraction methods are urgently needed for better information collection. Empirical mode decomposition (EMD) is an effective tool for nonstationary signal analysis, which has been widely applied in rolling element bearings and gearbox fault diagnosis. It has great advantage and adaptability in the mechanical fault diagnosis and feature extraction. EMD is a new time-frequency signal analysis method proposed by the scientist of National Aeronautics and Space Administration (NASA) Huang et al. in recent years [20]. This method has been broadly investigated by many researchers since it was provided. It has been applied in different areas for fault diagnosis. Parey et al. used EMD statistical method to detect incipient fault of the gears [21]. Loutridis applied the instantaneous energy density and EMD to monitor and diagnose the gear fault [22]. Liu et al. detected the gear incipient fault with EMD and they found that the result was better than that of wavelet decomposition [23]. EMD can also be used in machine fault diagnosis based on concrete analysis of specific issues. The vibration of blade crack generates a characteristic frequency (CF) which can be modulated into blade passing frequency (BPF). Therefore, EMD can be applied to determine the CF of blade cracks despite of the noise interference in practical working centrifugal compressor. It is also similar to gearbox fault diagnosis problems.

In this paper, PP signals are used for blade working condition classification by using EMD. Experiments are carried on to verify the effectiveness of this method in a test-rig. To verify the effectiveness of this method, strain testing is also carried on for the blade crack analysis. The structure of this paper is as follows. Section 2 introduces the theory of feature extraction for blade crack classification. Section 3 presents the simulation signal analysis. Section 4

describes our experimental setup for blade crack monitoring. Section 5 demonstrates PP signal analysis for blade condition classification. Section 6 gives concluding remarks.

2. Theory and Method

2.1. Empirical Mode Decomposition. EMD is developed based on instantaneous frequency calculation. It has been considered a very useful tool for the analysis of nonstationary and nonlinear signals [20]. For an arbitrary time series $X(t)$, it can decompose the original into many narrow-band components, each component known as intrinsic mode functions. An intrinsic mode function is used to convert it into a practically useful instantaneous frequency. The intrinsic mode function satisfies two conditions: (1) in the whole range of a data set, the number of the extreme must be equal to the number of the zero crossing points or the difference between them must be one; (2) at any given time, the mean value of the local positive extreme is equal to that of the local negative extreme. An arbitrary nonstationary and nonlinear signal can be decomposed into a series of components satisfied with the intrinsic mode function by using the local wave decomposition method. It is a sifting process and can be written as

$$\begin{aligned} X(t) - C_1(t) &= r_1(t) \\ r_1(t) - C_2(t) &= r_2(t) \\ &\vdots \\ r_{n-1}(t) - C_n(t) &= r_n(t). \end{aligned} \tag{1}$$

The original data can be decomposed into an n-series of intrinsic mode components plus a residual component r_n. The residual can be either a variable or a constant. Thus, the original signal can be expressed as

$$X(t) = \sum_{i=1}^{n} c_i(t) + r_n(t). \tag{2}$$

After EMD, intrinsic mode function (IMF) can be obtained. FFT can be used on different IMFs analysis for CF determination. EMD can be looked as a filter on feature determination. Therefore, it is helpful to obtain the CF.

2.2. Blade Characteristic. In general, centrifugal compressor casing vibration and radiation noise are closely related to blade BPF and its harmonics. It is also generated by the interference between rotor and stator during blade high speed rotation. BPF has high energy in the pressure frequency spectrum. It is the main source of centrifugal compressor noise. Its value can be determined by shaft speed multiplying the number of blade. BPF can be calculated by

$$\mathrm{BPF} = \frac{\mathrm{RPM}}{60} \times N, \tag{3}$$

where RPM is the shaft speed and N is the number of blades on the impeller.

BPF is the interference between rotator and stator. As BPF is a high frequency component, the low frequency components for blade nonorder vibration can be modulated to BPF during rotation. The modulation information will appear as the sideband frequency of the BPF. For unbalanced rotor conditions, SF will also be modulated to the BPF giving a sideband frequency around the BPF for unbalanced condition. Sideband frequency could be used to determine the modulated CF. The sideband frequency produced for blade cracks is different from SF. It can be used to warn blade crack. It does not mean that there is a blade with cracks if SF is the sideband frequency for BPF. It is also difficult to classify CF just according to the spectrum for the incipient crack as the magnitude of the blade vibration is weak compared with the amplitude of BPF. Therefore, effective feature extraction is urgently needed for blade crack analysis.

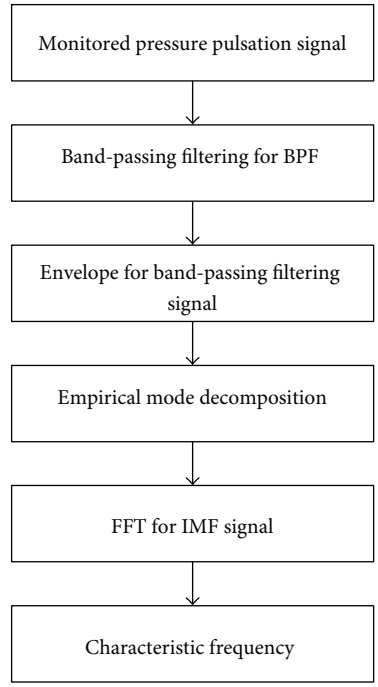


Figure 2: Flowchart for blade crack classification using PP signal.

2.3. Blade Crack Characteristic Frequency Determination. The steps for CF determination are shown in Figure 2. Firstly, PP is monitored based on the best suitable position according to blade crack classification. This is also a key step to determine the crack information because the sensor location has a direct effect on classification accuracy. Secondly, band-pass filter is applied on signal analysis. Envelope analysis is used to filter BPF signal. Then, IMFs can be obtained by using EMD. Fast Fourier transform is used on different IMFs. In the end, CF for blade crack can be obtained. Blade strain is used to verify the effectiveness of this method.

3. Simulation Signal Analysis

For a amplitude modulation signal sig(t), it can be expressed as

$$\mathrm{sig}(t) = A\left(1 + B\cos\left(2\pi F_e t\right)\right)\sin\left(2\pi F_c t\right), \tag{4}$$

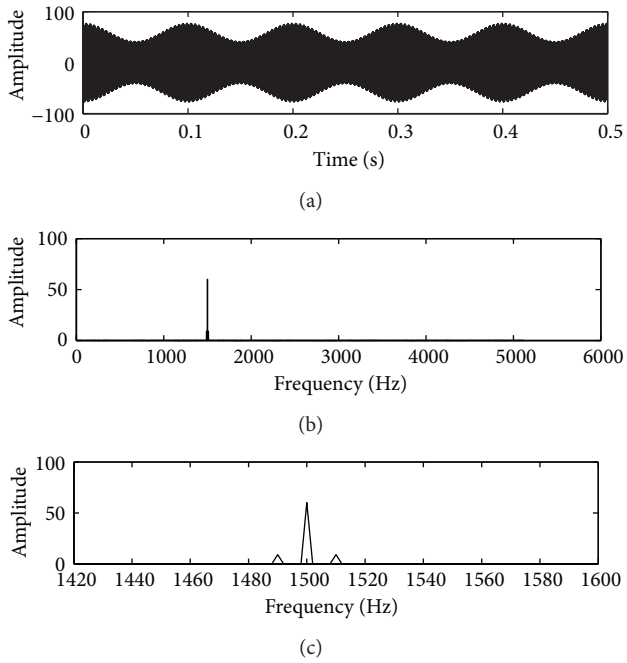


Figure 3: Signal demodulation analysis: (a) time domain signal for the simulation signal; (b) spectrum analysis for the simulation signal; (c) enlarged frequency area for the carrier frequency area.

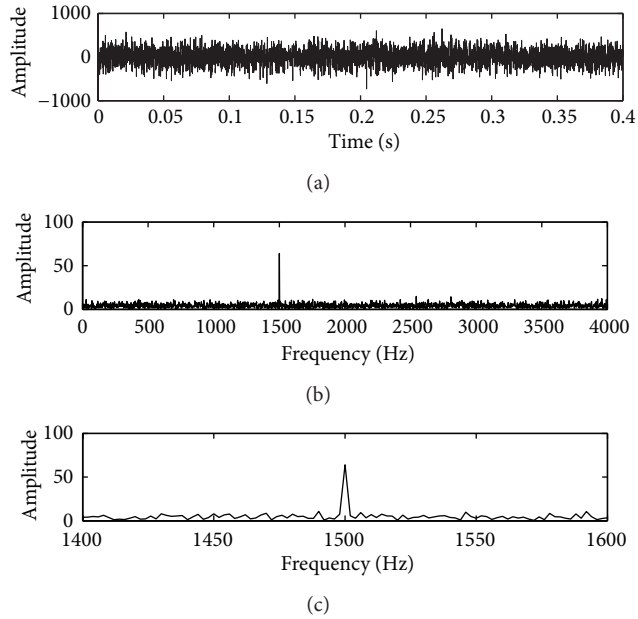


Figure 4: Frequency spectrum analysis: (a) time domain signal for the noise interference signal; (b) spectrum analysis for the noise interference signal; (c) enlarged frequency area for the carrier frequency area.

where, $F_c = 1500$ Hz, $F_e = 10$ Hz, $A = 60$, and $B = 0.3$. F_c and F_e correspond to carrier frequency and modulation frequency, respectively. The corresponding sampling frequency is 10,240 Hz for the simulation signal. Based on (4), an amplitude modulation signal can be obtained as shown in Figure 3(a). Fourier spectrum analysis is shown in Figure 3(b). The main frequency is 1500 Hz. The modulated frequency 10 Hz can be obtained by enlarging the frequency domain around the carrier 1500 Hz frequency shown in Figure 3(c). It is obvious for the sideband frequency around the carrier frequency if there is no noise interference in the signal.

Strong noise interference is added to the simulation signal as the characteristic information is usually overwhelmed by noise under practical working conditions. The obtained signal is shown in Figure 4(a). In the frequency spectrum analysis, there is clear broad frequency band noise effect shown in Figure 4(b). To determine the modulated signal, the enlargement for carrier frequency area in the spectrum is shown in Figure 4(c). Obviously, the enlarged frequency area is not clear due to the noise interference. The noise interference has an effect on the CF determination; therefore, it is difficult to classify the CF just according to sideband frequency spectrum analysis if there is strong noise interference.

Signal filter is used for the monitored signal around BPF. The filter band is 1400–1600 Hz. EMD is applied on the filter signal. IMFs can be obtained as shown in Figure 5. There is not any clear modulated frequency 10 Hz for every IMFs as shown in Figure 6. Envelope method is applied to the filter signal to filter BPF interference. EMD method is also applied on the envelope signal and IMFs can be obtained as shown in Figure 7. But there is clear modulated frequency 10 Hz as shown in Figure 8, the 5th IMFs based on EMD.

Based on above analysis, it can be convenient to determine the modulated frequency though there is strong noise interference. According to above analysis process, the blade nonorder vibration can be also monitored as it has the same property for simulation signal. Therefore, experimental verification is investigated in this research for blade nonorder vibration classification.

4. Experimental Test-Rig

4.1. Testing-Rig. To verify the effectiveness of this method, an experiment was carried on blade crack condition analysis by using the method based on PP signals analysis in a test-rig. The schematic diagram for the test-rig is shown in Figure 9. It contains an electric motor, fluid coupling, gearbox, and impeller. The impeller is a semiclosed one with 800 mm diameter. It is an experimental impeller for performance testing. By using fluid coupling, the rotating speed for impeller varies from 500 RPM to 9000 RPM. With the speed-up gearbox, the rotation speed of impeller can meet the designed one. The ratio between the driving and driven gears is 126/43 = 2.93. The experimental picture and hole in the diffuser for installing PP sensor are shown in Figure 10. The crack length during the experiment is 70 mm. PP, vibration, shaft speed sensors are installed to monitor the working process. There are 13 blades in this semiopen impeller. In this experiment, the speed of the impeller is 4500 RPM and 5000 RPM. The SF and BPF parameters are shown in Table 1.

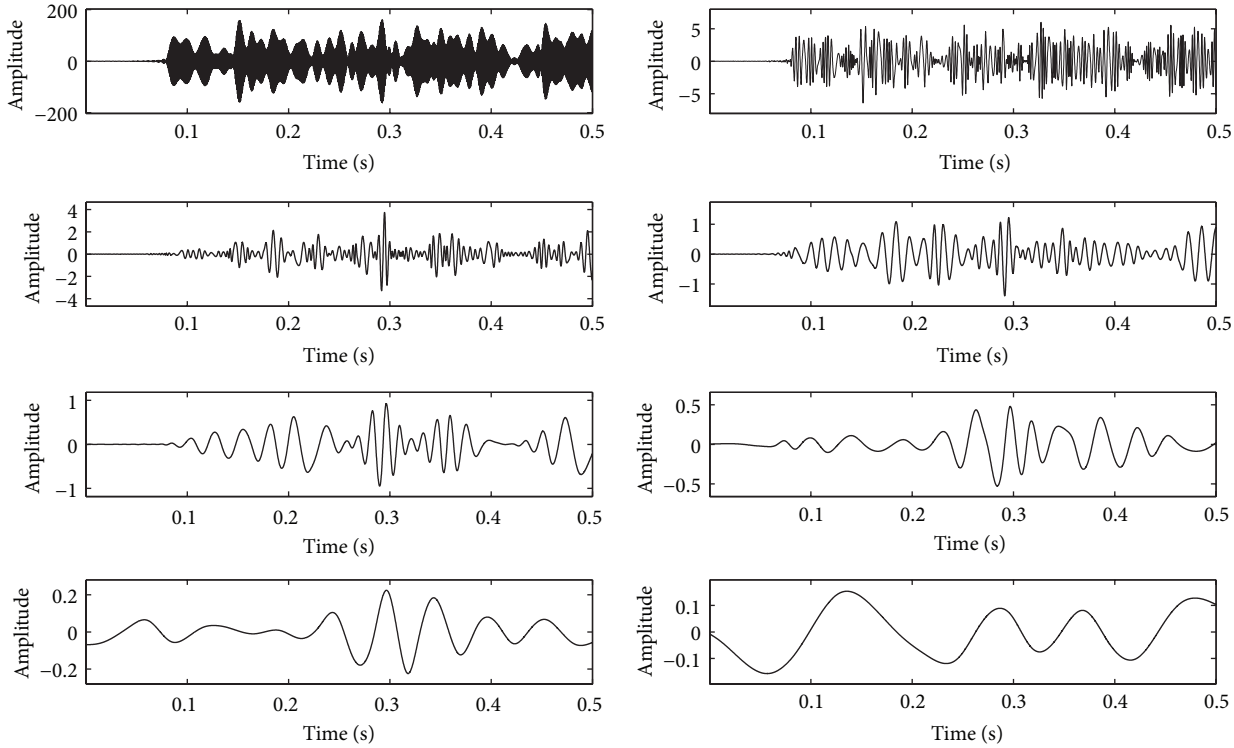


Figure 5: Time domain wave of IMFs based on EMD for the filter signal.

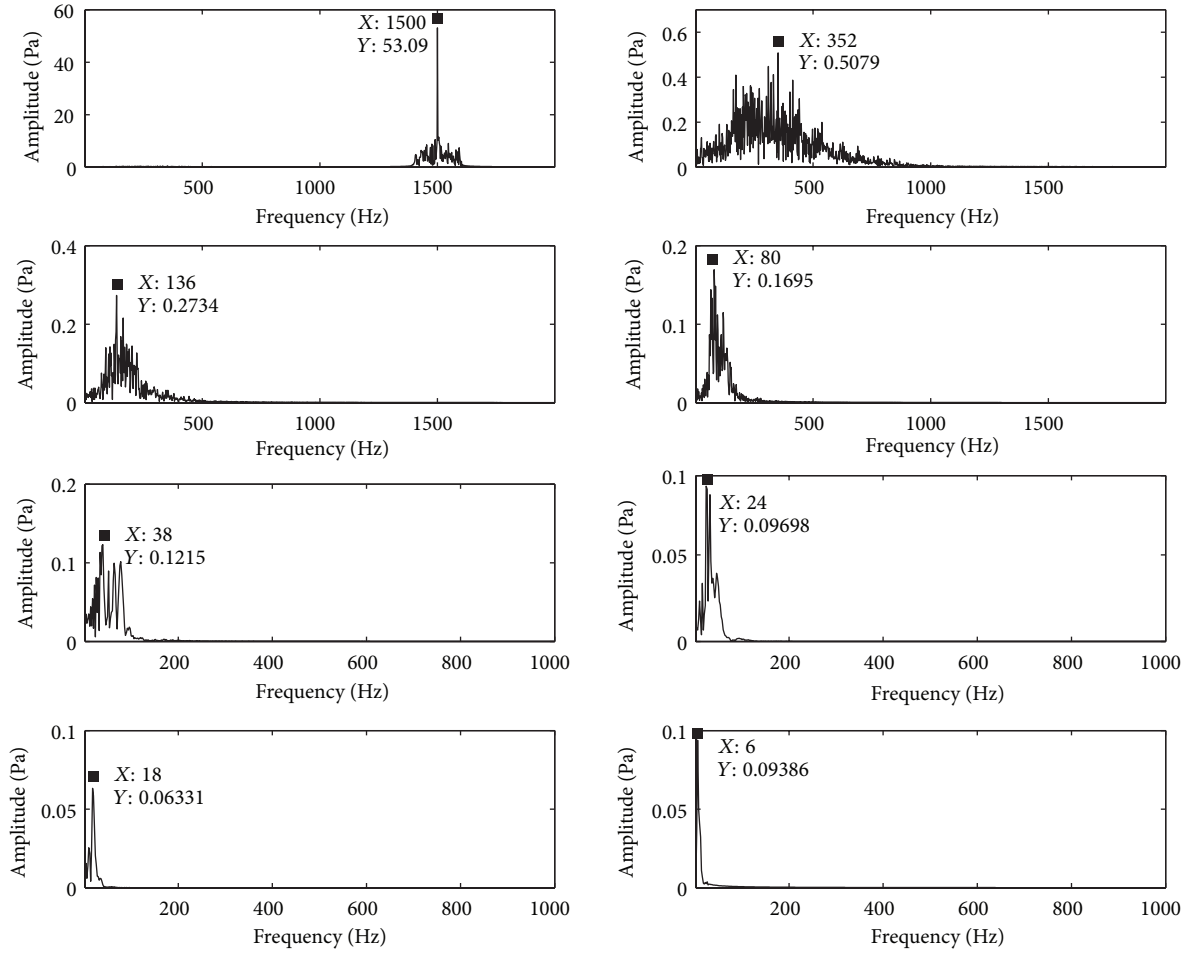


Figure 6: Spectrum of IMFs based on EMD for the filter signal.

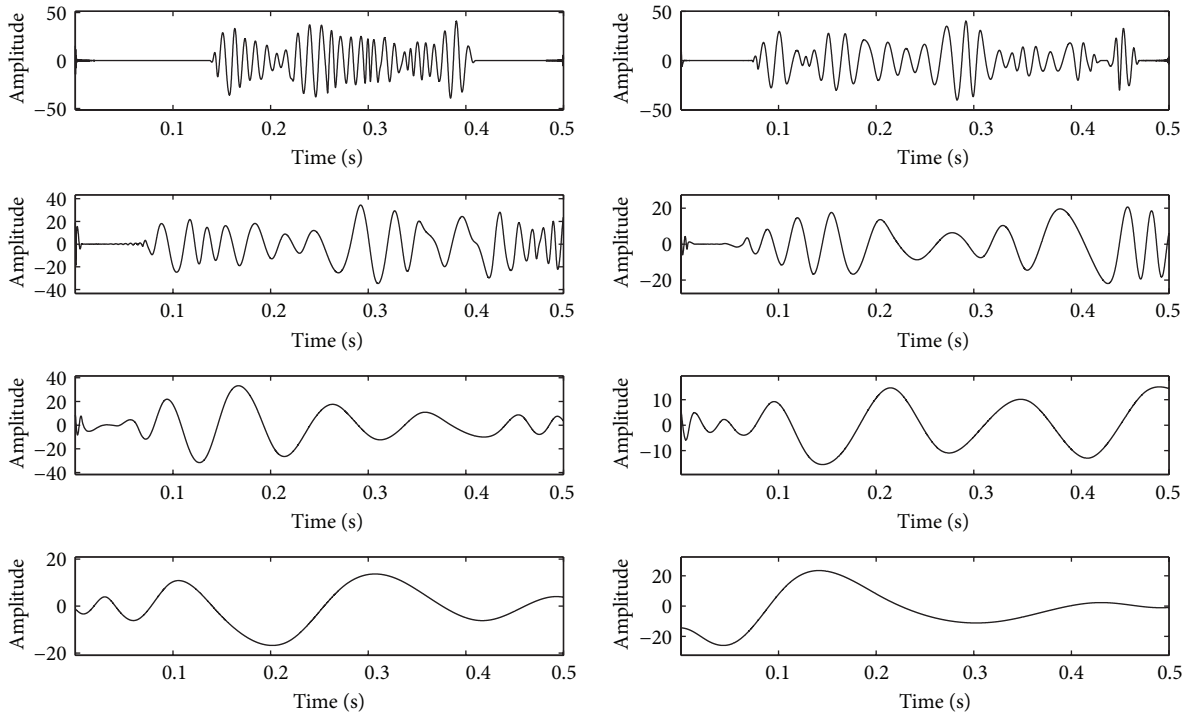


Figure 7: Time domain wave of IMFs based on EMD for the enveloped signal.

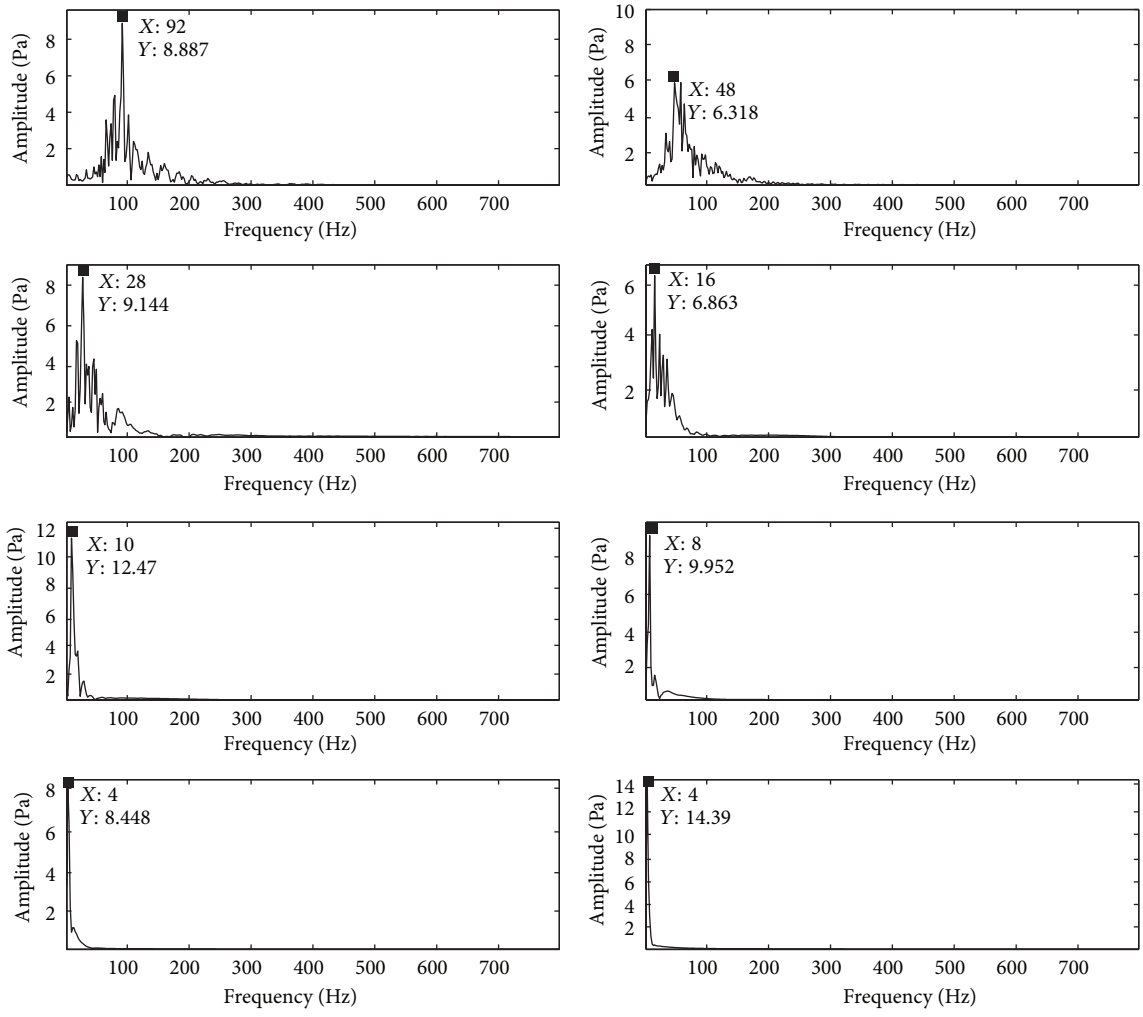


Figure 8: Spectrum of the IMFs for the envelope signal.

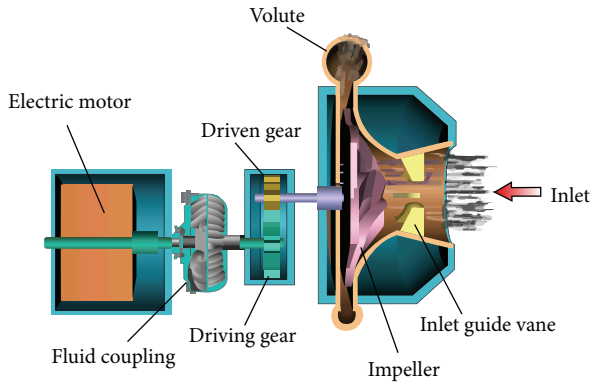


Figure 9: Experimental test-rig.

Figure 10: Pictures for test-rig.

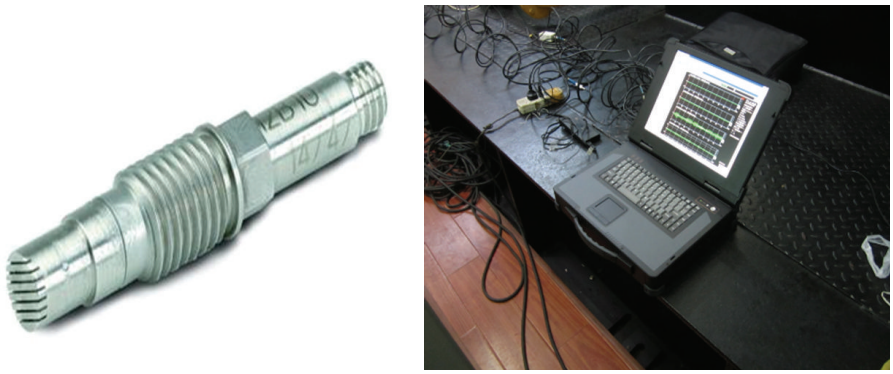


Figure 11: Pressure pulsation data acquisition system.

(a) Different strain sensor on the blade (b) Position for data acquisition module

Figure 12: Strain testing process.

(a) Data acquisition module (b) Data receiving module

Figure 13: Strain data acquisition system.

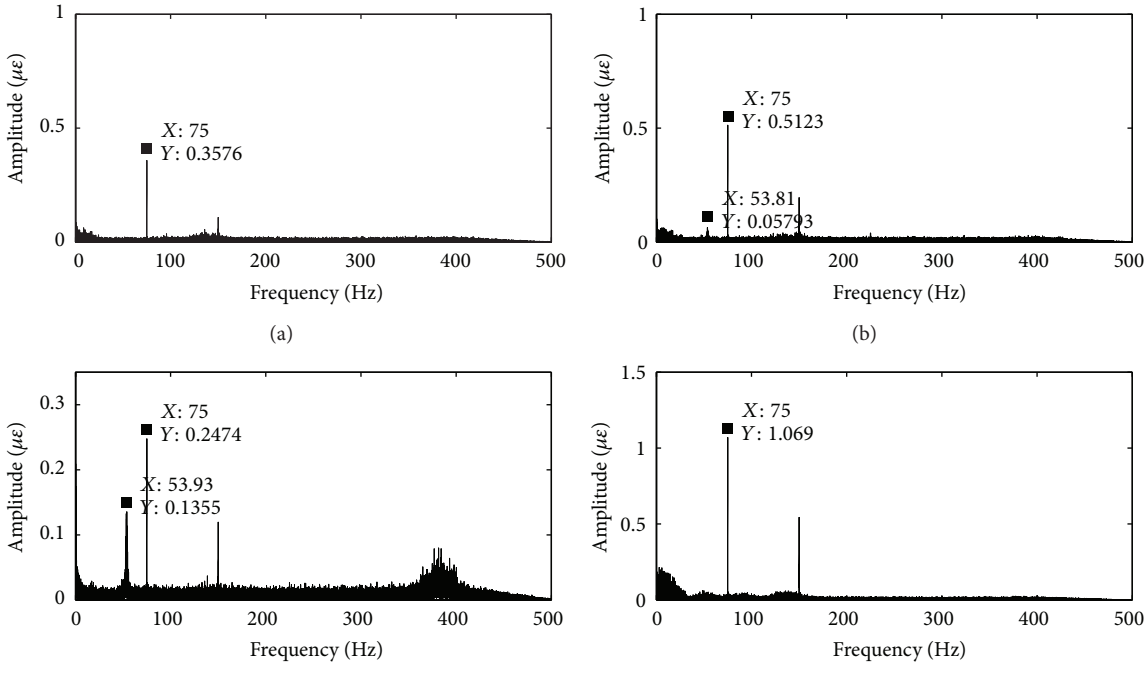


Figure 14: FFT Strain signal in 4500 RPM.

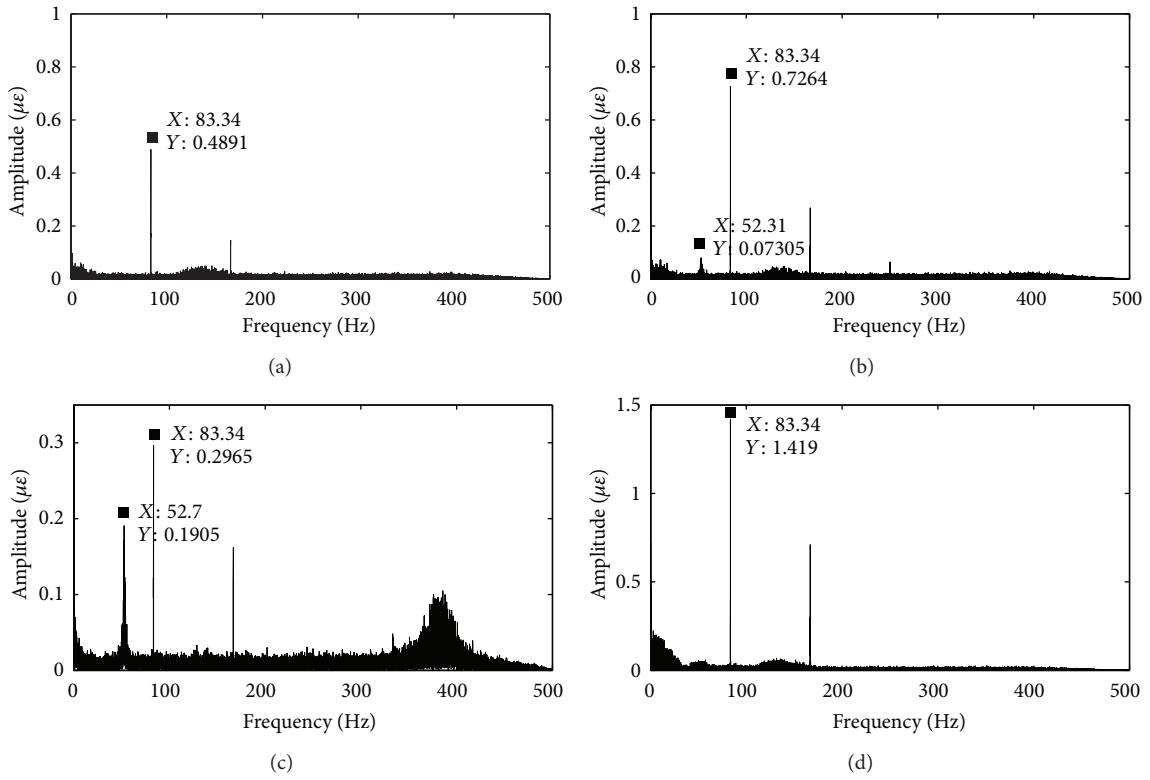
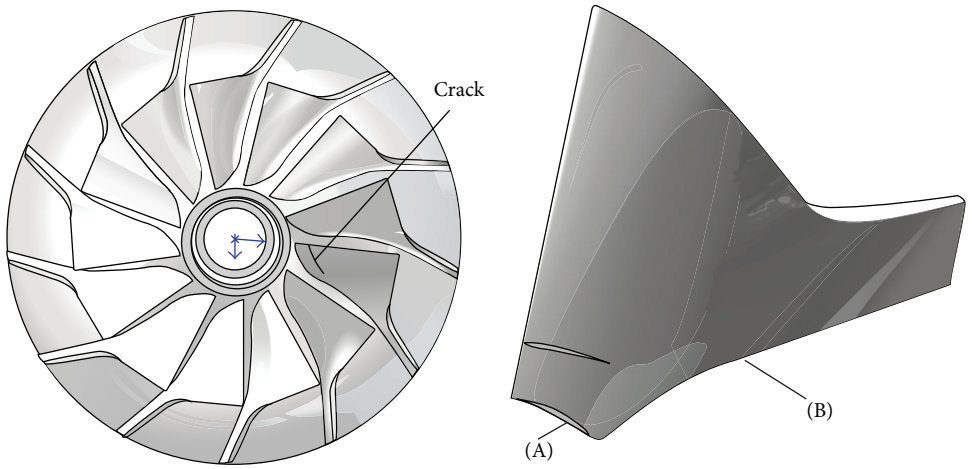


Figure 15: FFT Strain signal in 5000 RPM.



(a) Three models for impeller (b) A blade separated from impeller

Figure 16: Structure of impeller and blade.

Table 1: Characteristic parameters for φ800 test-rig.

Speed (RPM)	4500	5000
Shaft frequency (Hz)	75	83.3
Blade passing frequency (Hz)	975	1083

4.2. Data Acquisition. There are three PP sensors produced by PCB Piezotronics (New York, USA) to monitor the working process; it is shown in Figure 11. One is installed in the inlet pipe; the other two are installed near the diffuser in the holes shown as Figure 10. The sensitivities of the PP sensors are 0.7044 mV/Pa, 0.9845 mV/Pa, and 0.7336 mV/Pa. PP signal, vibration signal is gathered by the NI-4472 data acquisition card. It is an 8-channel synchronous data gathering system. It is also shown in Figure 11.

The blade strain test is carried out by this research to verify the appearance of the fault frequency. The blade vibration will lead to the strain changes on the surface of

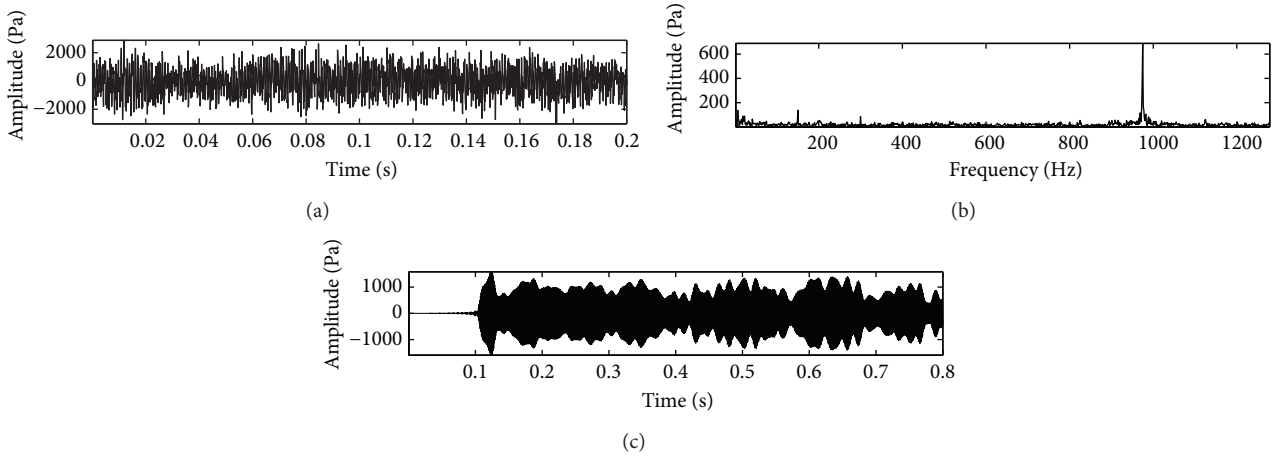


Figure 17: Time and frequency domain wave for the PP signal in 4500 RPM.

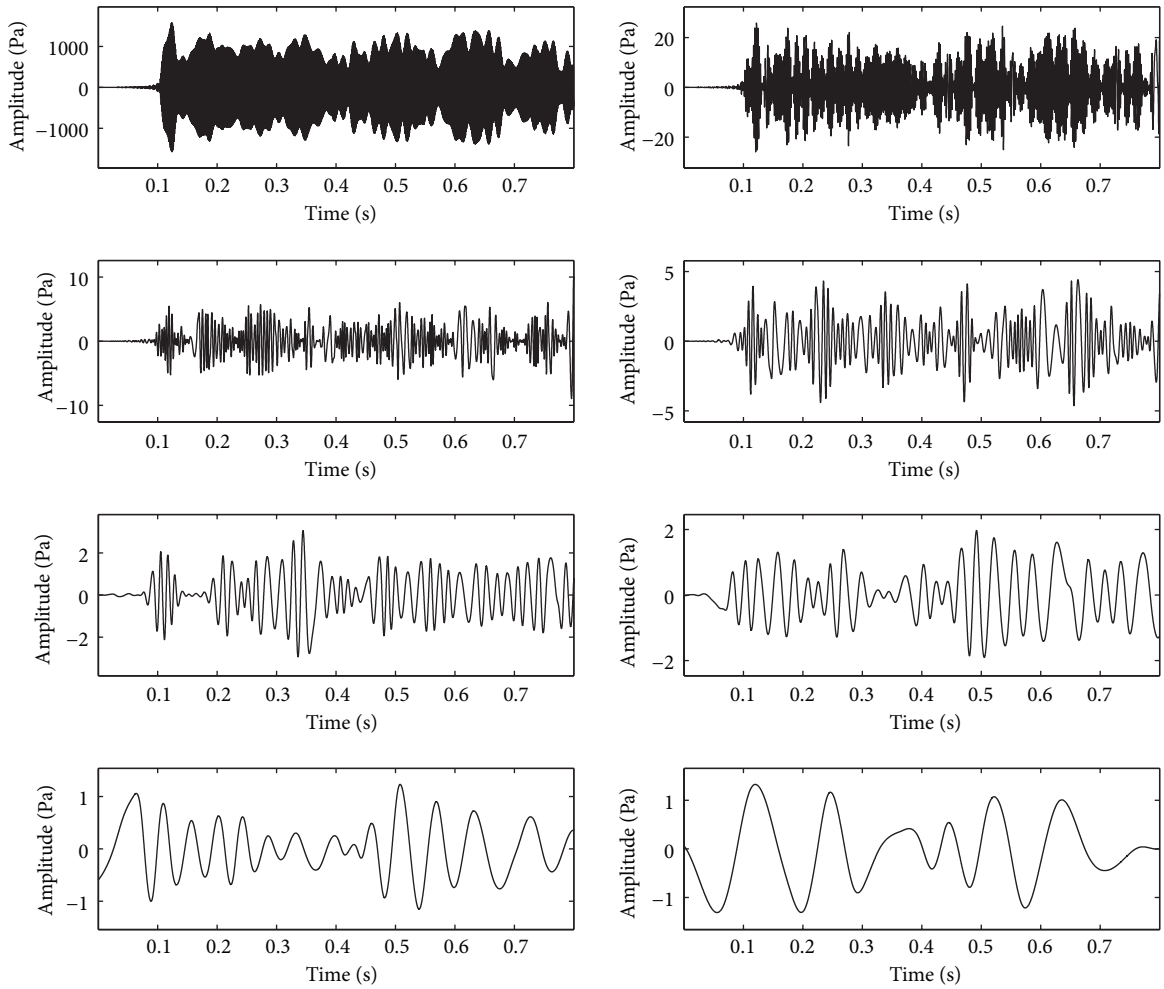


Figure 18: IMFs for the PP signal based on EMD in 4500 RPM.

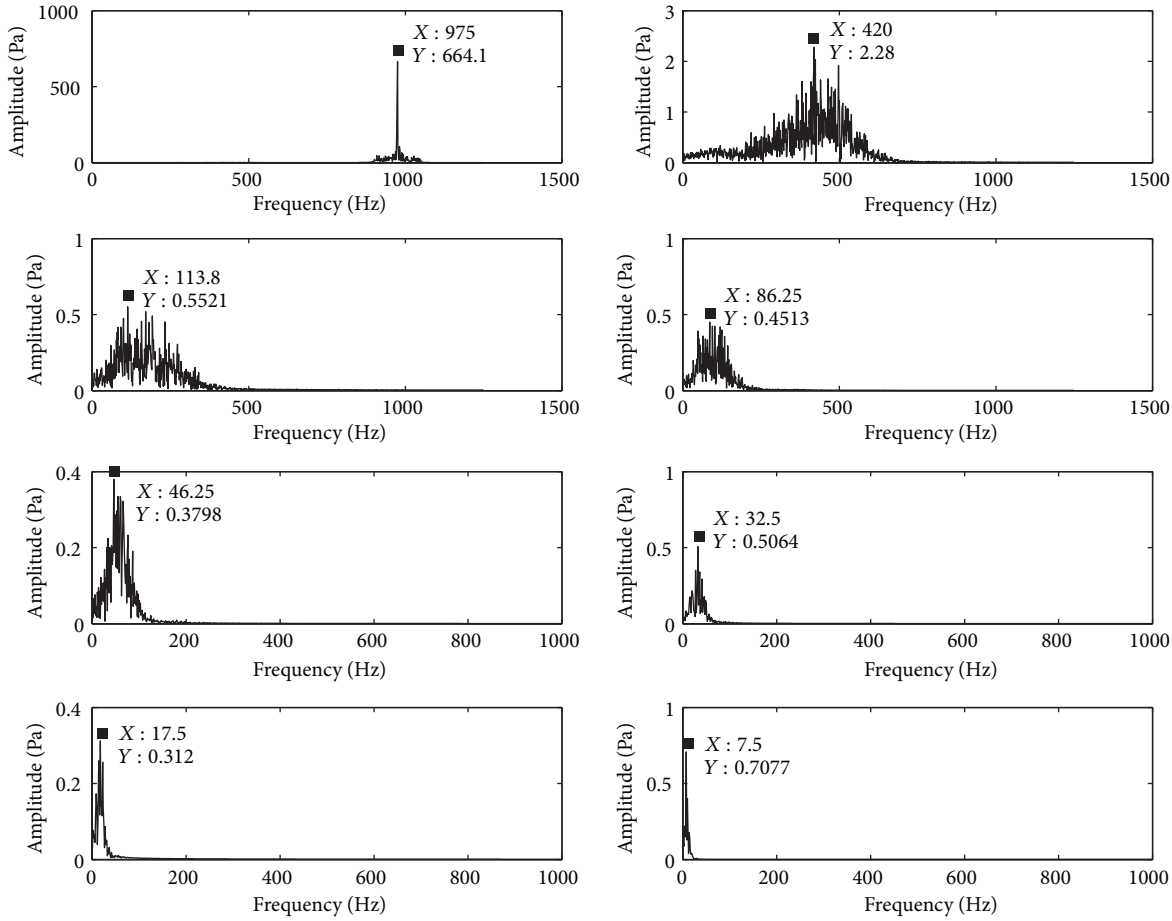


Figure 19: Spectrum for IMFs in 4500 RPM.

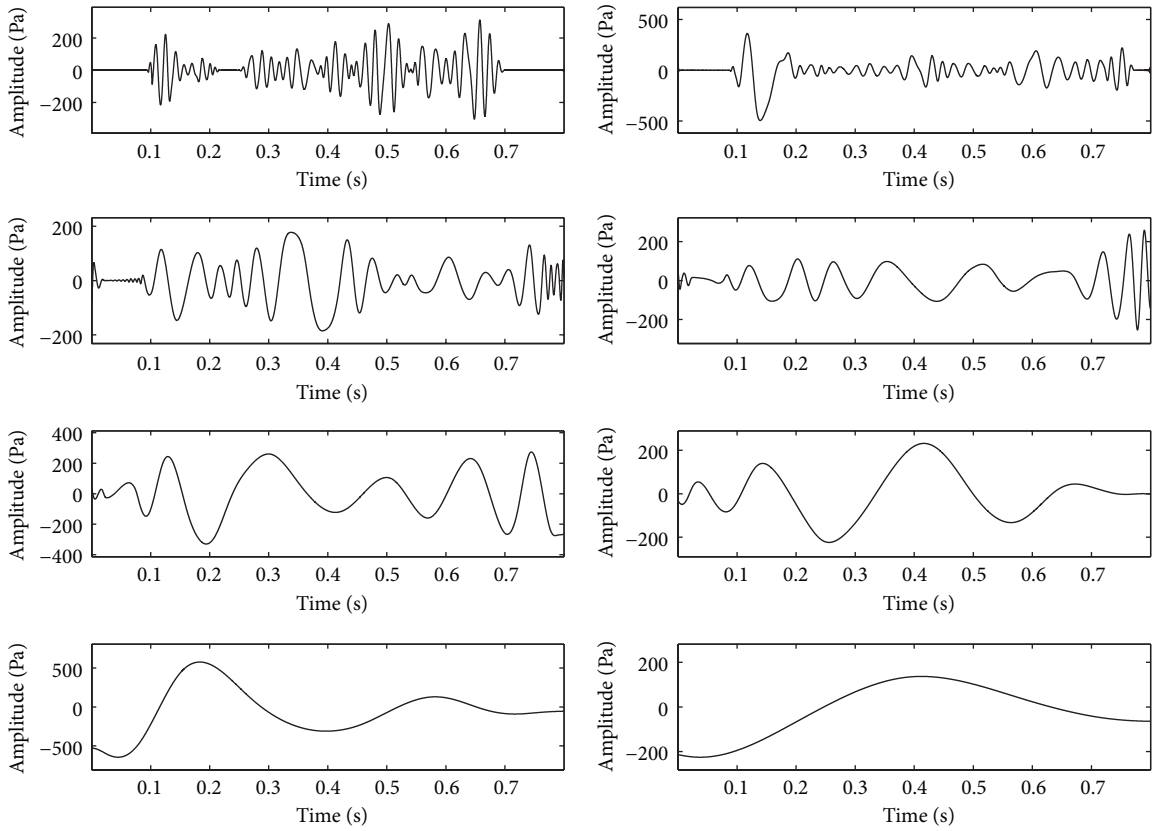


Figure 20: IMFs for the enveloped signal based on EMD in 4500 RPM.

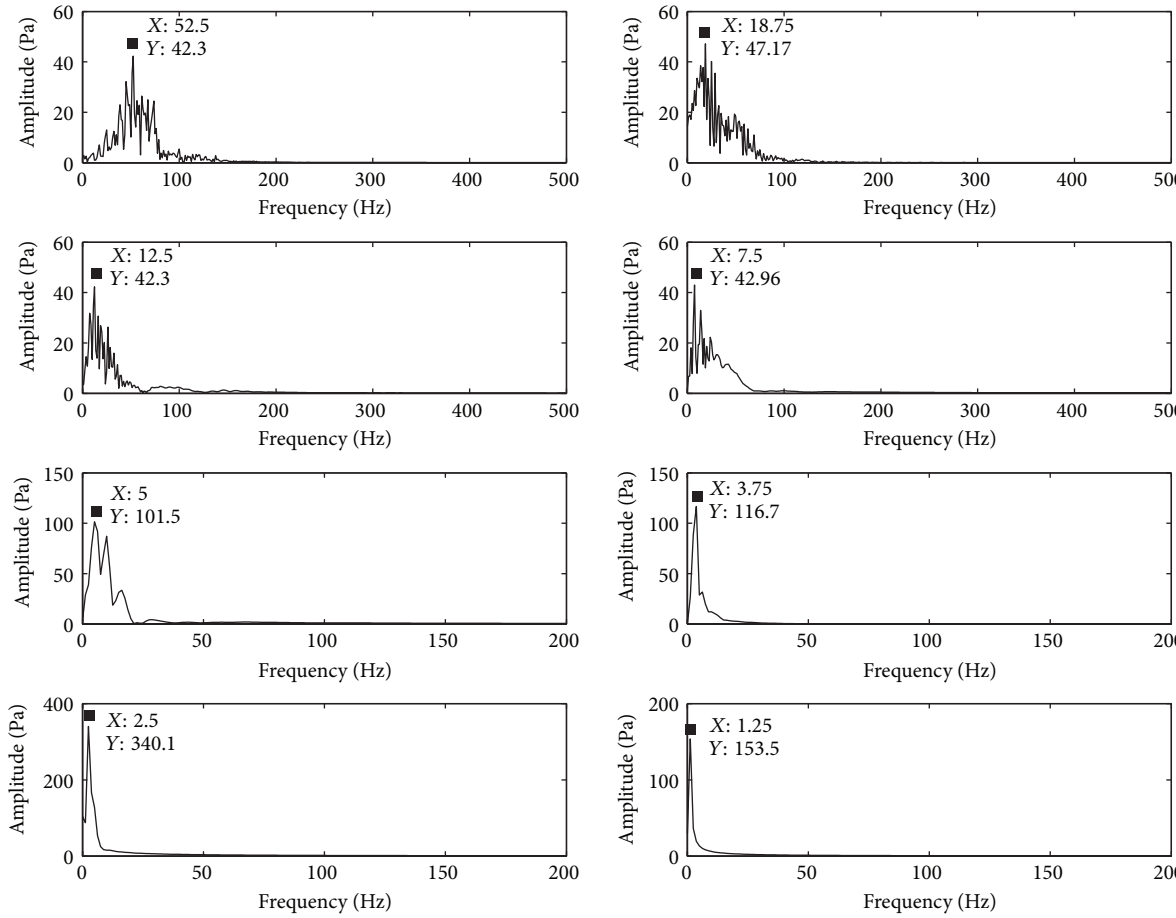


Figure 21: Spectrum for enveloped signal's IMFs in 4500 RPM.

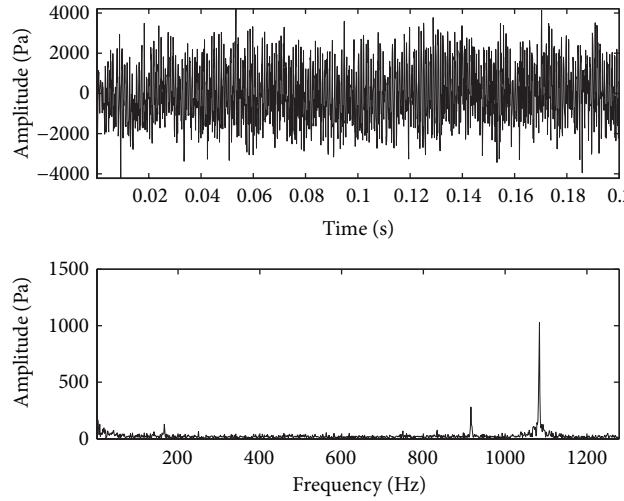


Figure 22: Time and frequency domain wave for the PP signal in 5000 RPM.

the blade, so the blade strain can reflect the blade vibration very well. When there are cracks on the blade, the blade vibration due to the change of the blade characteristics such as the stiffness and the blade stress vibration will change at the same time. So the stress test can be used to detect the blade crack. Due to real blade failure process, the location for the crack in selected near the hub shown in Figure 12. So the location for the crack is selected near the hub shown in Figure 12. To verify the appearance of the CF and its relationship with crack, the locations for strain gauge are shown in Figure 12(a). Points (a), (b), and (c) are on the crack blade. Point (d) is on a normal blade. The data acquisition

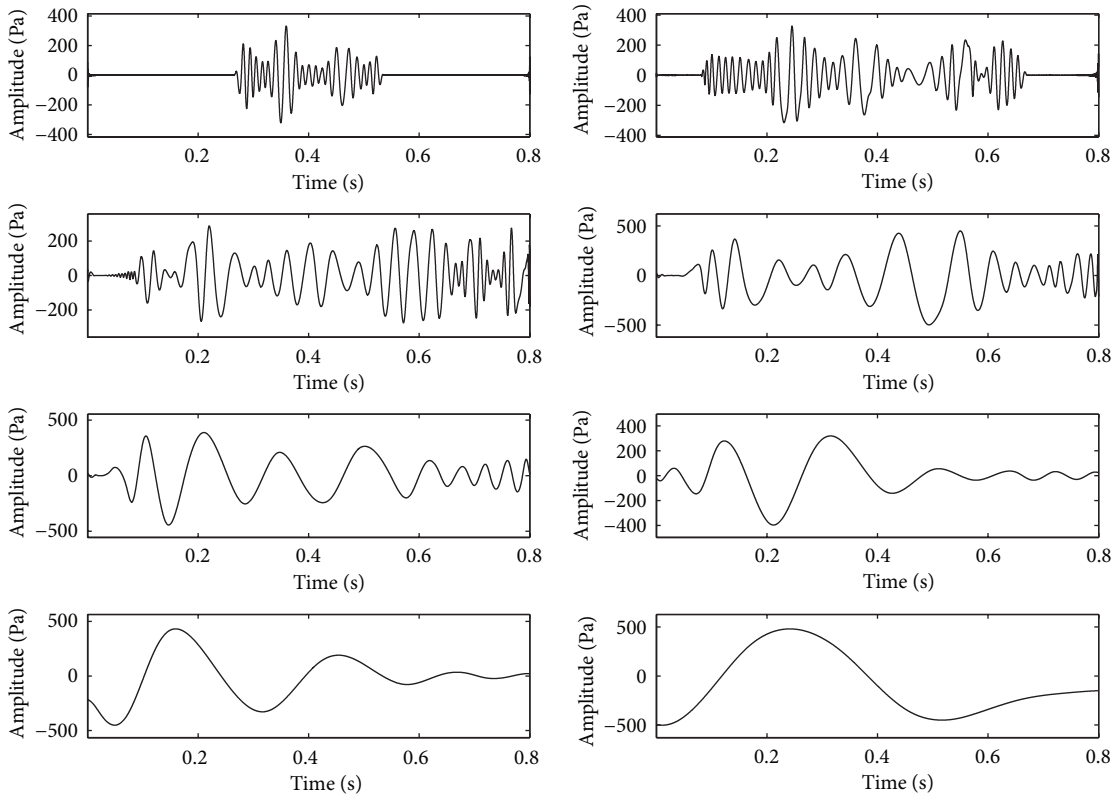


Figure 23: IMFs by EMD for the envelope signal in 5000 RPM.

module is shown in Figure 13(a) for launching data. It is also shown in Figure 12(b). Figure 13(b) presents the data receiving module. The data sampling frequency is 1024 Hz for strain signal. There are four channels to monitor the strain shown as Figure 12(a). It is important to determine the blade nonorder vibration as the reason of crack.

5. Data Analysis

5.1. Strain Signal Analysis. The frequency spectrum for the strain data with impeller speed 4500 RPM is shown in Figure 14. The SF of the impeller is 75 Hz. It is also clear because of the unbalance. The frequency 53 Hz is shown in the spectrum for point (b) and point (c). There is not the CF information for point (d) shown as Figure 14(d) because it is a normal blade. Point (c) is near the crack. It is clearer than point (b). It can be concluded that 53 Hz is the CF for blade vibration.

The same analysis is also carried on the impeller speed in 5000 RPM shown as Figure 15. Based on the above analysis process, the CF for blade nonorder vibration is 52.7 Hz. The CF is almost the same as speed in 4500 RPM. Therefore, it can be concluded that 53 Hz is the CF for the crack. It is a nonorder vibration for the blade and the reason of crack. There is not CF for the normal blade. Strain analysis can help us to determine the CF for blade crack. As it is not convenient in real working condition for strain monitoring, feature extraction is important to obtain the CF from other monitored signals.

5.2. Pressure Pulsation Signal Analysis. PP signal is used to detect blade nonorder vibration information as for the crack. Compared the total length of the blade, the crack is very small (the diameter for the blade is 800 mm) shown in Figure 16. Sides A and B in Figure 16(b) are together with impeller. They are not separated from impeller. It is just for clear demonstration with Figure 16(b). The impeller is manufacturing with whole milling process. At the same time, the averaging thickness of the blade is 10 mm to keep the stiffness of blade. Therefore, the information is weak for blade crack. It is the reason that it is difficult to determine the blade information. It can be just found when there is blade fracture. The crack information will be modulated to BPF as mentioned above although it is weak. There is not any modulated information in time domain shown in Figure 17(a). It is clear for BPF in the spectrum analysis shown in Figure 17(b). But it is not clear for the modulated frequency as the noise interference and nonorder vibration is very weak. It is impossible to obtain the modulated frequency. Therefore, the signal filter is investigated. The filter frequency band is 900–1050 Hz. The filter signal is shown in Figure 17(c). The time domain and the frequency domain spectrum are shown in Figures 18 and 19, respectively. But there is not any information about the CF.

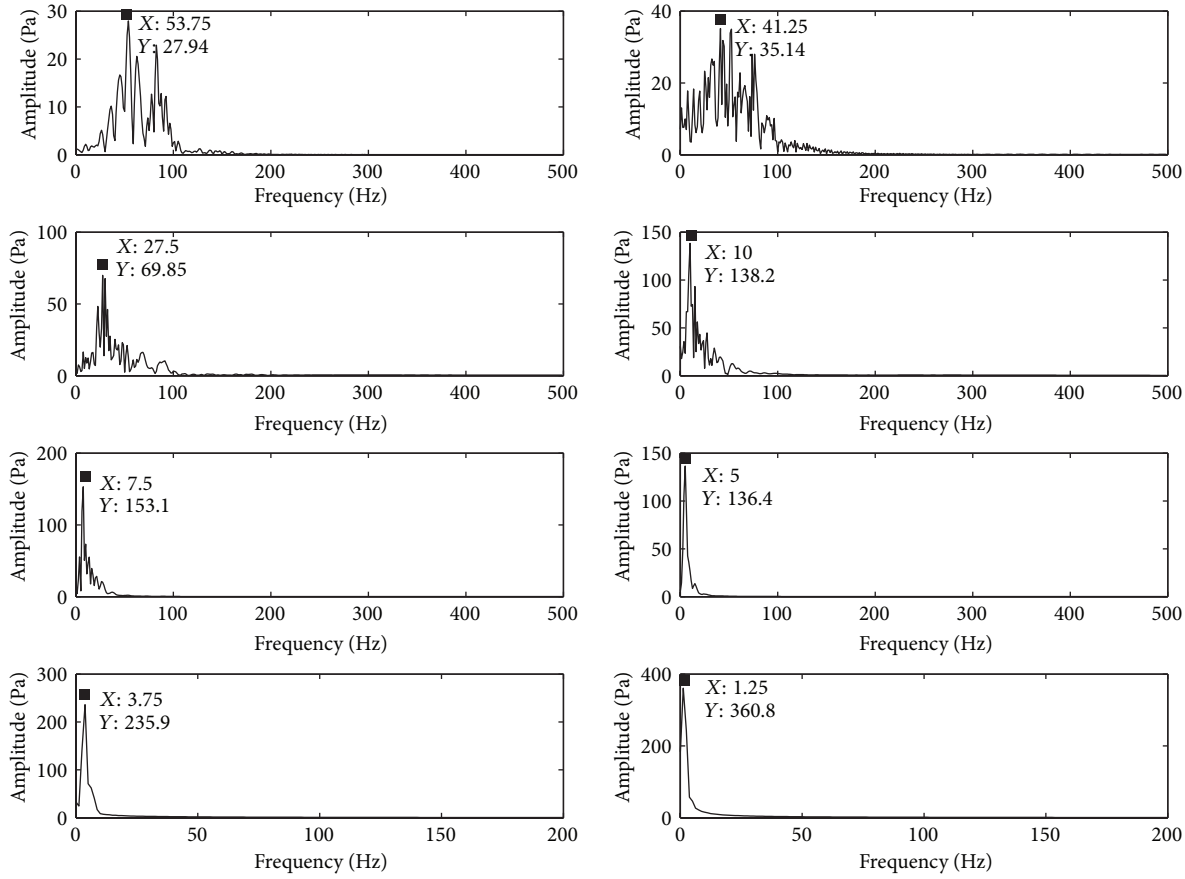


Figure 24: Time and frequency domain wave analysis for the different IMFs in 5000 RPM.

Envelope is used on the filtered signal. Then, EMD is used on envelope signal analysis. IMFs can be obtained shown in Figure 20. IMFs frequency spectrum analysis can clearly demonstrate the modulated frequency 53 Hz shown in Figure 21. Therefore, this method can be used to classify the crack CF for blade.

It is also with same result for 5000 RPM. Time and frequency analysis waveform for PP signal is shown in Figure 22. It is also difficult to recognize the CF. Therefore, signal filter is carried on. The filter frequency band is from 990 Hz to 1165 Hz. Then, IMFs based on EMD for envelope signal are shown in Figure 23. IMFs spectrum analysis is shown in Figure 24. It is clear for the modulated frequency 53 Hz. It has the same result with 4500 RPM. It can be verified that this method can effectively recognize the modulated nonorder vibration signal. It also demonstrates that this method can be used on feature frequency determination.

6. Conclusions

In this research, PP signals are used for blade crack condition monitoring and classification. The realization of this method is demonstrated in detail. Experiments on an industrial centrifugal compressor with a cracked blade were carried out to verify the effectiveness of this method. CF of blade crack information can be obtained by using EMD and spectrum analysis to obtain the modulated frequency. Strain signal is also investigated to monitor the crack CF. It is verified that crack characteristics can be determined by using PP signal. This research puts forward a method on how to determine the blade crack CF. Further investigations will also be carried on how to apply this method on real working condition blade crack classification. It will be helpful for blade crack early warning.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The work was supported by the Natural Science Foundation of China under Grant no. 51175057 and the Fundamental Research Funds for the Central Universities under Grant no. DUT14ZD204.

References

[1] Z. Y. Huang and X. F. Wang, *Turbine Compressor*, Chemical Industry Press, Beijing, China, 2004.

[2] A. Kammerer, *Experimental Research into Resonant Vibration of Centrifugal Compressor Blades*, Swiss Federal Institute of Technology, Zürich, Switzerland, 2009.

[3] B. Wen, X. Wu, Q. Ding et al., *Theory and Experiment of Nonlinear Dynamics for Rotating Machinery with Faults*, Science Press, Beijing, China, 2004.

[4] M. Baumgartner, F. Kameier, and J. Hourmouziadis, *Non-Engine Order Blade Vibration in a High Pressure Compressor*, ISABE, Melbourne, Australia, 1995.

[5] Y. G. Lei, J. Lin, Z. He, and D. Kong, "A method based on multi-sensor data fusion for fault detection of planetary gearboxes," *Sensors*, vol. 12, no. 2, pp. 2005–2017, 2012.

[6] N. J. Lourenço, M. L. A. Graça, L. A. L. Franco, and O. M. M. Silva, "Fatigue failure of a compressor blade," *Engineering Failure Analysis*, vol. 15, no. 8, pp. 1150–1154, 2008.

[7] A. Kermanpur, H. S. Amin, S. Ziaei-Rad, N. Nourbakhshnia, and M. Mosaddeghfar, "Failure analysis of Ti6Al4V gas turbine compressor blades," *Engineering Failure Analysis*, vol. 15, no. 8, pp. 1052–1064, 2008.

[8] F. L. Eisinger and R. E. Sullivan, "Vibration fatigue of centrifugal fan impeller due to Structural-Acoustic coupling and its prevention: a case study," *Journal of Pressure Vessel Technology*, vol. 129, no. 4, pp. 771–774, 2007.

[9] N. Roy and R. Ganguli, "Helicopter rotor blade frequency evolution with damage growth and signal processing," *Journal of Sound and Vibration*, vol. 283, no. 3–5, pp. 821–851, 2005.

[10] K. Elbhbah and J. K. Sinha, "Vibration-based condition monitoring of rotating machines using a machine composite spectrum," *Journal of Sound and Vibration*, vol. 332, no. 11, pp. 2831–2845, 2013.

[11] K. Saravanan and A. S. Sekhar, "Crack detection in a rotor by operational deflection shape and kurtosis using laser vibrometer measurements," *Journal of Vibration and Control*, vol. 19, no. 8, pp. 1227–1239, 2013.

[12] X. B. Liu, J. G. Lin, and Y. Wang, "Research on fault identification of blade crack of fan based on wavelet-packet analysis," *Machine Tool & Hydraulics*, vol. 35, no. 9, pp. 241–243, 2007.

[13] A. Rama Rao and B. K. Dutta, "Vibration analysis for detecting failure of compressor blade," *Engineering Failure Analysis*, vol. 25, pp. 211–218, 2012.

[14] H. H. Yang, H. Hou, X. Y. Zeng, and J. C. Sun, "Fault diagnosis for fan based on auditory spectrum feature of sound signal," *Chinese Journal of Scientific Instrument*, vol. 30, no. 1, pp. 175–179, 2009.

[15] L. Witek, "Experimental crack propagation and failure analysis of the first stage compressor blade subjected to vibration," *Engineering Failure Analysis*, vol. 16, no. 7, pp. 2163–2170, 2009.

[16] Y. Qu, C. Z. Chen, X. G. Zhao, and B. Zhou, "Wavelet scalogram identification for crack feature of wind turbine blade," *Journal of Shenyang University of Technology*, vol. 34, no. 1, pp. 22–47, 2012.

[17] X. Wang, H. Mao, H. Hu, and Z. Zhang, "Crack localization in hydraulic turbine blades based on kernel independent component analysis and wavelet neural network," *International Journal of Computational Intelligence Systems*, vol. 6, no. 6, pp. 1116–1124, 2013.

[18] B. C. Zhou, C. Zhang, and M. Yu, "Research on dynamic propagating characteristics of wind turbine blade's cracks," *China Mechanical Engineering*, vol. 24, no. 8, pp. 1108–1113, 2013.

[19] E. Egusquiza, C. Valero, X. Huang, E. Jou, A. Guardo, and C. Rodriguez, "Failure investigation of a large pump-turbine runner," *Engineering Failure Analysis*, vol. 23, pp. 27–34, 2012.

[20] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *The Royal Society of London. Proceedings. Series A. Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[21] A. Parey, M. El Badaoui, F. Guillet, and N. Tandon, "Dynamic modelling of spur gear pair and application of empirical mode decomposition-based statistical analysis for early detection of localized tooth defect," *Journal of Sound and Vibration*, vol. 294, no. 3, pp. 547–561, 2006.

[22] S. J. Loutridis, "Instantaneous energy density as a feature for gear fault detection," *Mechanical Systems and Signal Processing*, vol. 20, no. 5, pp. 1239–1253, 2006.

[23] B. Liu, S. Riemenschneider, and Y. Xu, "Gearbox fault diagnosis using empirical mode decomposition and Hilbert spectrum," *Mechanical Systems and Signal Processing*, vol. 20, no. 3, pp. 718–734, 2006.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 124968, 8 pages
http://dx.doi.org/10.1155/2014/124968

Research Article

A PCA and ELM Based Adaptive Method for Channel Equalization in MFL Inspection

Zhenning Wu,[1] Huaguang Zhang,[1] Jinhai Liu,[1] Zongjie Qiu,[2] and Mo Zhao[1]

[1] *School of Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110004, China*
[2] *CNOOC (China) Co., Ltd., Beijing 100010, China*

Correspondence should be addressed to Zhenning Wu; wuzn2003@hotmail.com

Received 28 April 2014; Revised 30 June 2014; Accepted 14 July 2014; Published 12 August 2014

Academic Editor: Ruqiang Yan

Magnetic flux leakage (MFL) as an efficient method for pipeline flaw detection plays important role in pipeline safety. This nondestructive test technique assesses the health of the buried pipeline. The signal is gathered by an array of hall-effect sensors disposed at the magnetic neutral plane of a pair of permanent magnet in the pipeline inspection gauge (PIG) clinging to the inner surface of the pipe wall. The magnetic flux measured by the sensors reflects the health condition of the pipe. The signal is influenced by not only the condition of the pipe, but also by the lift-off value of the sensors and various properties of electronic component. The consistency of the position of the sensors is almost never satisfied and each sensor measures differently. In this paper, a new scheme of channel equalization is proposed for MFL signal in order to correct sensor misalignments, which eventually improves accuracy of defect characterization. The algorithm proposed in this paper is adaptive to the effects of error on the disposition of the sensor due to manufacturing imperfections and movements of the sensors. The algorithm is tested by data acquired from an experimental pipeline. The results show the effectiveness of the proposed algorithm.

1. Introduction

MFL is a widely used nondestructive testing (NDT) methods for pipeline inspection. The inspection machine is usually called PIG. A PIG using MFL consists of several pairs of strong permanent magnets which magnetize the pipe along the axial direction of the pipe. Each pair of the strong permanent along with the yoke iron and the hall sensors and also brush is called a carrier. The carrier with the pipe consists of a magnetic circuit. Details of the PIG can be found at some famous inspection companies website, ROSEN, PII, and so forth. A lot of research work has been done to analyze the signal of MFL. Carvalho et al. [1], Christen and Bergamini [2], and Xiang and Tso [3] purposed neural network based methods to detect flaws. Mukhopadhyay and Srivastava [4] proposed wavelet based technique to denoise the signal of the MFL inspection. Mukherjee et al. proposed wavelet based inverse mapping system [5]. Kathirmani et al. [6] using PCA [7] and wavelet [8–11] technique to compress data of the MFL signal.

But there is still one problem that needs to be studied. The sensing arrangement, that is, each sensor, its mechanical support, and the underlying electronics for acquiring the magnetic leakage flux data, commonly referred to as a channel, suffers mismatch among each other. A lot of factors can cause channel-to-channel mismatch, including the lift-off value between the pipeline, and position of hall and coil sensors and the various properties of electronic component. Other factors influence the factors mentioned above can also impact the output of the signal, such as the difference of the sensors location caused by assembly, the shake when the detector running in the pipeline, and so forth. All these factors make the output of the signal different even under same testing condition and testing object. This will lower the capacity of the detector especially during the post processing of the signal. Such kind of mismatch may also exist in other multisensor data processing. Commonly, it can be equalized by using adaptive techniques. An adaptive channel equalization algorithm to deal with the problem of channel-to-channel mismatch of MFL signals is given in

[12]. In [12], they assume that at least one sensor out of the sensor array is ideal and can be used as a reference. For implementation, this assumption imposes serious limitations on the performance of post processing algorithm as tolerance and misalignment of an individual sensor is not deterministic and needs to be accounted for in a stochastic framework for choice of a clear cut candidate qualifying as a reference channel. To solve the problem, Mukherjee et al. [13] gives an adaptive method which does not need to choose a reference channel. In [13], the reference channel required for channel equalization is replaced with the baseline estimation. The baseline estimation reflects the background leakage flux. But the baseline estimation is not available under a defect or feature. It is estimated by first order forward or back prediction of the neighboring MFL data.

In this paper, a new adaptive channel equalization algorithm to minimize channel-to-channel mismatch is proposed for MFL signals. In contrast to [12] our algorithm does not need to choose any reference channel. Because the ideal reference channel is not easy to get and the character of each channel may have little difference, equalizing the channels adaptively with no reference channel is reasonable. The signal of all channel is learned by a neural network. The training data set is selected as a clean pipe (almost no flaw), which reflects the character of the pipe. And distinguished from [13], we mainly focus on the signal around and including the flaw. In [13], the signal around the flaw is only estimated by first order forward or backward prediction of the neighboring MFL data, which may cause distortion of the flaw signal. As flaw evaluation needs the exact shape of the signal around and including the flaw, our algorithm has more advantage. A PCA based flaw detection algorithm is given in this paper to find the location of the flaw signal. A median corrected algorithm is also given from the engineering point of view. The simulation results show that the algorithm proposed in this paper is efficient.

The paper is organized as follows. In Section 2, an ELM based method is given to dynamically compensate each channel, and also with a PCA based statistic method to separate normal and flaw signal and extract signal characters. The details of our algorithm proposed in this paper are stated too. A median corrected algorithm is given, and simulation results are shown in Section 3. Section 4 concludes the paper indicating major achievements and future scope of this work.

2. PCA and ELM Based Channel Equalization

2.1. Channel Equalization Using ELM Neural Networks. Neural networks are very efficient and popular tool to do regression and classification. The advantages of the neural networks are mainly two points: one is that the model of the data does not need to be known, the other is its high capability to deal with nonlinear problems. There exist many types of neural networks; however, feedforward neural networks may be one of the most popular neural networks. The feedforward neural network usually consists of one input layer receiving the stimulin from external environments, one or multihidden layers, and one output layer sending the network output to external environments. Widely used neural networks include backpropagation (BP) neural network [14], radial basis function (RBF) neural network [15], and support vector machine (SVM) [16]. Three main approaches are usually used to train feedforward networks including gradient-descent based method (e.g., BP neural networks), least-square based method (e.g., RBF network learning) and standard optimization method based method (e.g., SVM). Different from traditional learning algorithms, ELM [17] tends to reach not only the smallest training error but also the smallest norm of output weights. According to the neural network theory, for feedforward neural networks, smaller training error results in smaller norm of weights and better generalization performance. Since the hidden layer needs not be tuned in ELM and the hidden layer parameters can be fixed, the output weights can then be resolved using the least-square method. The model of channel equalization using ELM is given as Figure 1.

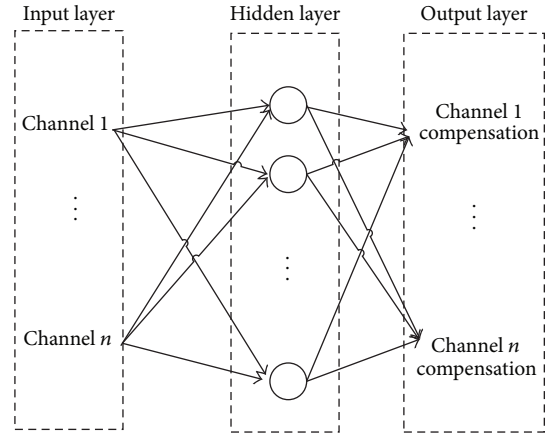


Figure 1: Neural network model for channel equalization.

The output function of single-hidden layer feedforward networks (SLFNs) with L hidden nodes can be represented by

$$f(x) = \Sigma_{i=1}^{l} \beta_i g_i(x), \tag{1}$$

where $g_i(x)$ denotes the output function of the ith hidden node.

For N arbitrary distinct samples (x_i, t_i), SLFNs with L hidden nodes are mathematically modeled as

$$\Sigma_{i=1}^{l} \beta_i g_i\left(x_j\right) = y_j, \quad j = 1, \ldots, N. \tag{2}$$

That SLFNs can approximate these N samples with zero error means that

$$\Sigma_{j=1}^{l} \left\| y_j - t_j \right\| = 0. \tag{3}$$

The parameters in $g_i(x)$ can be trained according to (2).

This can be written as

$$G\beta = T, \tag{4}$$

where

$$G = \begin{bmatrix} G(x_1) \\ \vdots \\ G(x_N) \end{bmatrix} = \begin{bmatrix} g(w_1, b_1, x_1) & \cdots & g(w_l, b_l, x_1) \\ \vdots & \ddots & \vdots \\ g(w_1, b_1, x_N) & \cdots & g(w_l, b_l, x_N) \end{bmatrix}. \quad (5)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_l^T \end{bmatrix}_{l \times m}, \qquad T = \begin{bmatrix} T_1^T \\ \vdots \\ T_l^T \end{bmatrix}_{N \times m}.$$

It is proved in [17] that given any small positive value $e > 0$, activation function $g : R \to R$ which is infinitely differentiable in any interval and N arbitrary distinct samples (x_i, t_i), there exists $l \leq N$ such that for any $\{w_i, b_i\}_{i=1}^{l}$ (parameters need to be trained in $g_i(x)$) randomly generated from any intervals of $R^d \times R$ according to any continuous probability distribution, with probability one, $\|H_{N \times l}\beta_{l \times m} - T_{N \times m}\| < e$. And from the interpolation point of view the maximum number of hidden nodes required is not larger than the number of training samples. In fact, if $L = N$, the training errors can be zero.

Though the ELM has good regression ability, there is still one obverse problem that the results of the ELM rely on the training data set. But the flaw difference is from not only length and width, but also depth, which may cause the MFL signal variance. It is impossible to include all flaw signal in the training data set. A better way is to use a small training data set to train the ELM which can generate good compensation results. One solution is given in this paper in Section 2.2.

2.2. PCA Based Flaw Exclusion. To train the ELM and solve the problem mentioned in Section 2.1, one solution is given.

Because the property of pipe is learned using ELM, we can use the channel with no flaw to predict the channel with flaw when flaw is detected. By using the predicted result to substitute the channels which detect flaw, the compensation result can be obtained using ELM. And then, using the ELM result, the compensation is given to each channel. This avoids training ELM with every type of flaw. To exclude the flaw, a PCA based algorithm is stated as follows, which detects which channels and which sampling points detest flaw signal.

PCA [18] as an efficient statistical learning algorithm is useful to deal with multivariable problems. The PIG has n sensors surrounding the vessel. Consider one sampling of all sensors as $x = (x_1, x_2, \ldots, x_n)^T$. The linear transform of the sensors result can be written as

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = a_1^T x \\ y_2 &= a_{11}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = a_2^T x \\ &\vdots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = a_n^T x \end{aligned} \quad (6)$$

y_1 is called the ith principle component. The computation steps is as follows.

First, m samples from each sensor are collected and written as a matrix $X \in R^{m \times n}$. The matrix is scaled to zero mean, and in addition to unit variance.

The second step is to compute the singular values. The covariance matrix is computed as

$$\Sigma \cong \frac{1}{m-1} X^T X. \quad (7)$$

An SVD (singular value decomposition) is used to compute the principal components and the associated singular vectors as

$$\begin{aligned} \frac{1}{m-1} X^T X &= T \Lambda T^T \\ \Lambda &= \begin{bmatrix} \Lambda_{\text{pc}} & 0 \\ 0 & \Lambda_{\text{res}} \end{bmatrix}, \end{aligned} \quad (8)$$

where

$$\begin{aligned} \Lambda_{\text{pc}} &= \text{diag}\left(\sigma_1^2, \ldots, \sigma_l^2\right) \in R^{l \times l}, \quad \sigma_1 \geq \cdots \sigma_l \geq \cdots \sigma_n \\ \Lambda_{\text{res}} &= \text{diag}\left(\sigma_{l+1}^2, \ldots, \sigma_n^2\right) \in R^{(n-l) \times (n-l)} \\ T &= \left[T_{\text{pc}} \;\; T_{\text{res}}\right] \in R^{(n \times n)} \\ T_{\text{pc}} &\in R^{n \times l}, \qquad T_{\text{res}} \in R^{n \times (n-l)} \\ TT^T &= T_{\text{pc}} T_{\text{pc}}^T + T_{\text{res}} T_{\text{res}}^T = I_{n \times n} \\ T_{\text{pc}} T_{\text{pc}}^T &\begin{bmatrix} T_{\text{pc}}^T \\ T_{\text{res}}^T \end{bmatrix} \left[T_{\text{pc}} \;\; T_{\text{res}}\right] \end{aligned} \quad (9)$$

l is the number satisfying

$$\frac{\Sigma_{i=1}^{l} \sigma_i^2}{\Sigma_{i=1}^{n} \sigma_i^2} \geq \alpha, \quad 0 \leq \alpha \leq 1 \quad (10)$$

$\sigma_i^2 / \Sigma_{i=1}^{n} \sigma_i^2 \geq \alpha$ is called the significance level. And (10) is called the cumulated significance level, which shows how much the first l principle components can reflect data X. The Hotelling T^2 statistic is used to detect fault

$$T_i^2 = \Sigma_{j=1}^{l} \frac{y_{ij}^2}{\lambda_j}. \quad (11)$$

With a set threshold, flaw signal can be separated from normal signal. Suppose there are only two sensors. Take 500 sampling. The result is shown in Figure 2. The linear transform of y is also shown in this figure. Using the Hotelling T^2 statistic, which is shown as formula (11), the flaw data can be detected.

2.3. Details of PCA and ELM Based Algorithm for Channel Equalization. The data gathered using a PIG can be described as $X \in R^{m \times n}$, where m is the sampling number and n is the number of sensors. The algorithm proposed in this paper treats data with two points of view, one is from the sampling view and the other is from the sensor view.

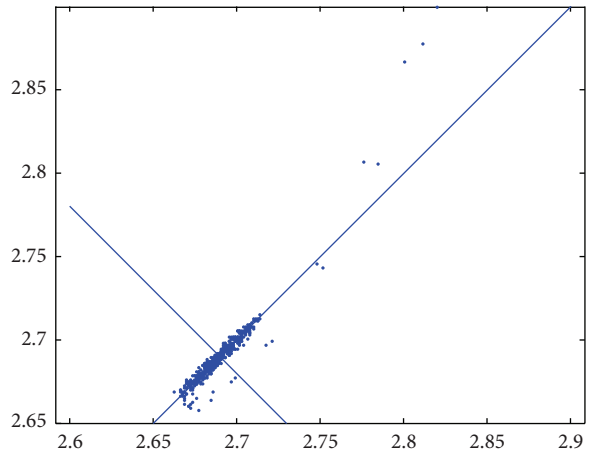


Figure 2: PCA analysis of 500 sampling using 2 sensors.

From the sampling view, the PCA algorithm is used to detect flaw, which determines at which sampling points the flaw locates. For example, take m sampling, using PCA, flaw locates at $[k, k+\Delta]$, where $1 \le k \le m$ and $1 \le k+\Delta \le m$ can be detected with a preset T^2 statistic. The normal part is tested using a trained ELM neutral network to remove channel mismatches, which is called ELM 1. And the flaw part is treated from the sensor point of view. The normal part is added to the training set dynamically in order to learn more character of this part of pipe.

From the sensor view, the flaw data is also treated using PCA to determine which sensors detect flaw. For example, only sensors s to $s+\Delta s$ detect flaw, where $1 \le s \le s+\Delta s \le n$. The normal channel is used as input data, using ELM trained with channels of normal from training data set, the flaw part of signal can be forecasted as normal part. This step is to revert the flaw part to its normal condition, in order to determine how much each channel needs to be compensated with ELM 1. And using ELM 1, the channel mismatch of the flaw part can be removed. It is clear that each flaw needs an ELM neural network to compute the compensation, which is called the ELM p in Figure 3.

The flow chart of the algorithm is shown in Figure 3 with steps and details stated as follows.

Step 1. An ELM is trained using training data set to learn the character of the normal condition of the pipe for channel equalization. The target is each channels compensation. The ELM trained is marked as ELM 1.

Step 2. Using PCA with $x = (x_1, x_2, \ldots, x_n)^T$ represents sensors to detect flaw which is stated in Section 2.2 with a threshold. The T^2 statistic up over the threshold denotes the flaw signal. The signal will be separated into normal parts and flaw parts in Step 3.

Step 3. The signal of flaw detected from Step 2 is tested using PCA with $x = (x_1, x_2, \ldots, x_m)^T$ represents sampling points. A threshold is also set automatically. The T^2 statistic up over the threshold denotes the channels which detects flaw. It makes the flaw signal detected from Step 2 separated into two parts, the signal of normal channels and the signal channel of flaw.

Step 4. Signal of normal parts acquired from Step 2 is tested using ELM 1 trained in Step 1.

Step 5. For the flaw signal detected in Steps 2 and 3, another ELM is trained with training data set. The normal channels are used as inputs of the ELM, and the flaw channels are used as output of the ELM. For example, the flaw is detected at sampling point from t to $t+\Delta t$, and channels from s to $s+\Delta s$. The training data set from channel 1 to $s-1$ and $s+\Delta s+1$ to n is used to train this neural network. The target of the ELM is set as the training data set with channels from s to $s+\Delta s$. Each flaw has an ELM. The ELM is marked as ELMp with p representing the number of flaw.

Step 6. The flaw signal from t to $t+\Delta t$, and channels from s to $s+\Delta s$ is replaced temporarily by the test result with ELMp.

Step 7. The signal segment in Step 6 is tested using ELM 1. The output is compensated to the original signal from t to $t+\Delta t$ and channels from s to $s+\Delta s$.

Step 8. Update the training data set with normal data acquired in Step 2.

3. Experiment Results

The PIG used to collect data in this paper consists of 15 carriers with 5 axial sensors on each carrier. An 8-inch seamless steel pipe with length of about 14 meters is used to do this experiment. 9 exterior flaws were made on this pipe. The pipeline in our experiment is connected with two flanges and one weld. The sampling is controlled by an odometer wheel with the sampling frequency of 1 sampling per 2 mm. Several experiments were done with different load angel of the PIG.

In order to train our algorithm, some signal of MFL of normal condition pipeline with no flaw is needed. The training data is obtained in two way. (i) The first way is to obtain from the original training data. One test data is selected as original training data with flaw signal excluded manually. (ii) The second way of getting the training data is generated from each test using algorithms stated in Sections 2.1 and 2.2. The algorithm used in this paper treats data batch by batch. One batch of data treated using the algorithm can separate this batch of data into normal data and flaw data. The normal part of data is added to the training data set in order to reinforce the training result. By adding new training data, the training set is updated. New character of the normal condition is studied. This means as the process goes, the result of the algorithm proposed in this paper gets better. To avoid the training data set getting too large, the length of the data set is set to a certain length. When length of the training data set grows up to its limits, the earlier training data will be erased and new training data will be added to replace the vacancy.

And in order to show the efficiency of our algorithm, the length of the original training data is reduced to only less than 1/5 of one test, though theoretically a bigger training

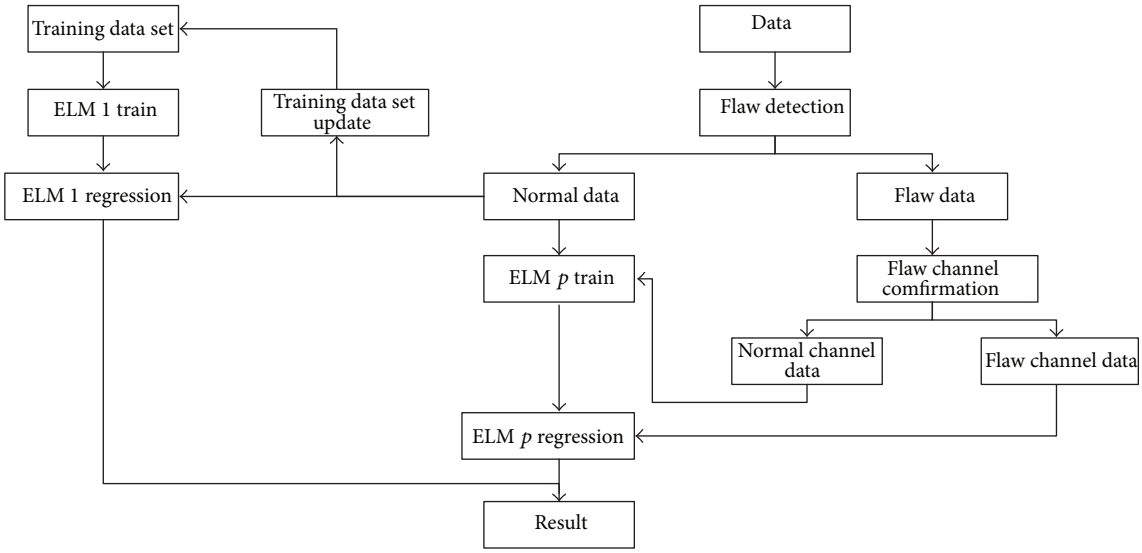


Figure 3: Flow chart of algorithm proposed.

data set will get better result. Also restricted to the experiment environment, it is not possible to build a pipeline with enough length of hundreds of meters or even miles. The results with less original training data also show that the algorithm proposed in this paper generates satisfying results.

And algorithm of median corrector is also adopted to compare with our results. Among many factors that influence channel-to-channel mismatch, the lift-off value and the difference of the baseline play the main role. The influence caused by difference of the lift-off value among sensors can be reduced by improving the assembly skills. And the influence of the baseline (zero output) of sensors can be reduced by calculating the average of sensors as baseline. The channel equalization can be computed as three steps. Assume X as data gathered by sensing a clean pipe (almost no flaw). First, calculate each channels average $\overline{x_i}$. Second, calculate x_{mean} the average of $\overline{x_i}$. Third, compensate $x_{\text{mean}} - \overline{x_i}$ to each channel. But as the MFL data is processed automatically, this method needs to be operated manually. And it is not that easy to find such steady state signal. To overcome these, we use median corrector to do rough channel equalization in engineering. Instead of calculating the average of each channel in the first step, $\overline{x_i}$ is each channel's median value. Other steps are same as the average method.

All data of one test is treated using the algorithm proposed in this paper. Figure 4 shows the PCA result of Step 2 in Section 2.3. The threshold is automatically generated with F distribution parameter of 95%. All the components and flaws are described in Figure 4. Details of raw signal of 3 flaws are shown to illustrate the following steps in Figure 5. The dashed line shows the flaw sampling point intervals detected using PCA as shown in Figure 5. By applying the sensor view of PCA stated as Step 3 in Section 2.3, the flaw signal channels are marked within the solid line in Figure 5. The sensor view PCA result is shown in Figure 6, with F distribution

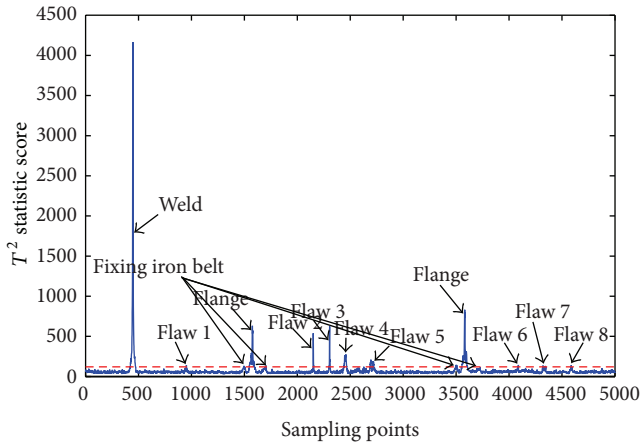


Figure 4: PCA flaw exclusion result of sampling view.

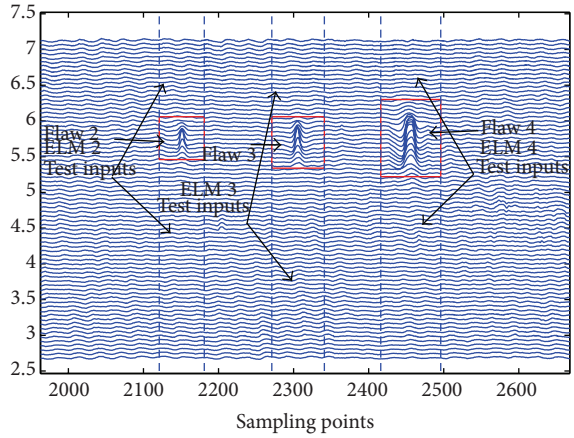


Figure 5: Detail of flaw signal.

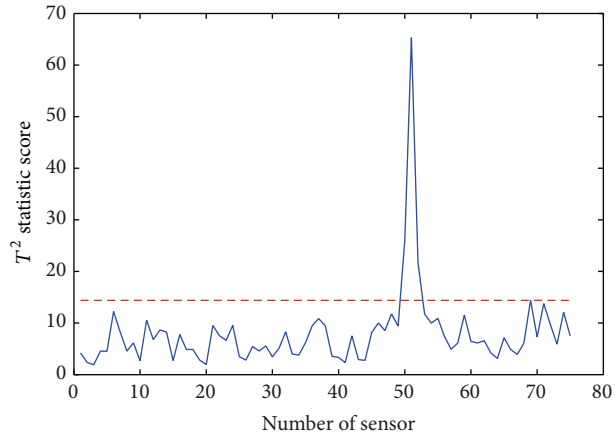


Figure 6: PCA flaw exclusion result of sensor view.

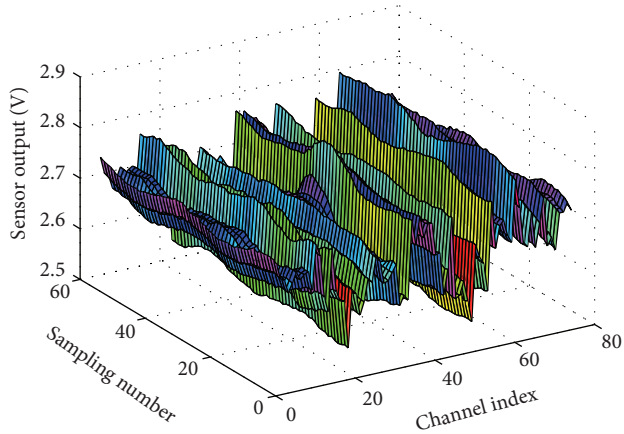


Figure 7: Raw data of flaw 1.

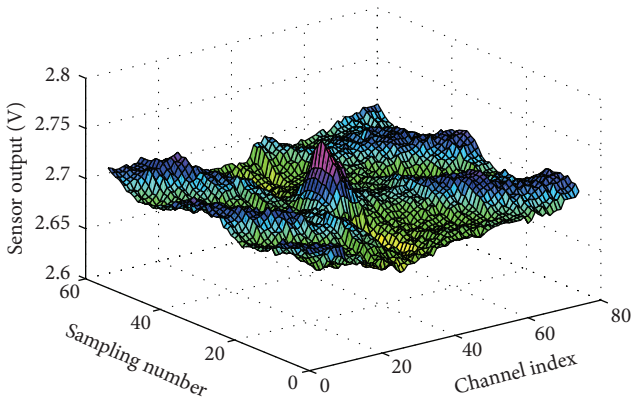


Figure 8: Median treated data of flaw 1.

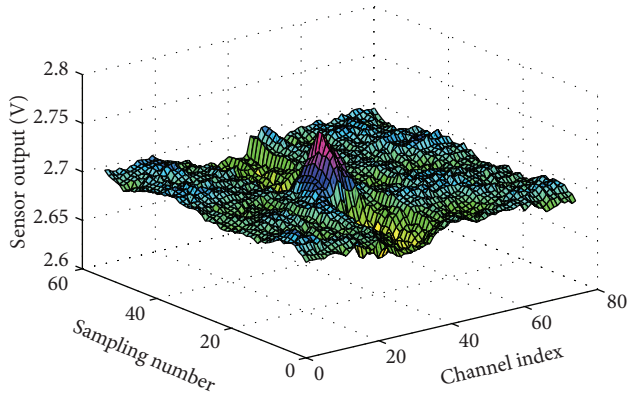


Figure 9: PCA and ELM based algorithm treated data of flaw 1.

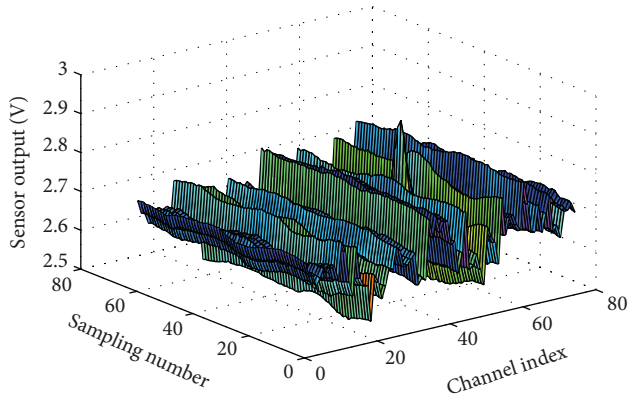


Figure 10: Raw data of flaw 2.

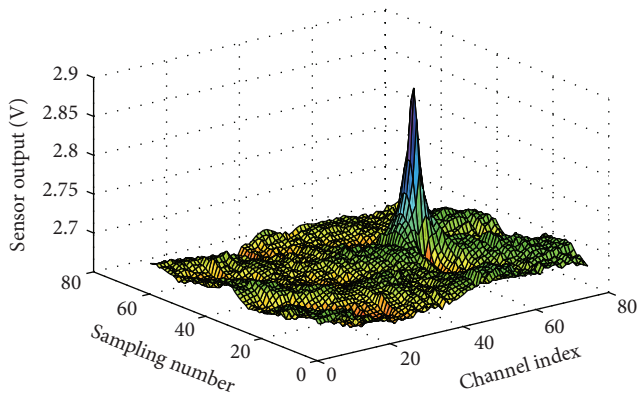


Figure 11: Median treated data of flaw 2.

parameter of 95%. The normal channels is used as input of ELM p, and test result is shown from Figures 7–12.

To show the result, the standard deviation (SD) and peak signal to noise ratio (PSNR) are adopted. The results are shown in Tables 1 and 2. Lower standard deviation reflects that the signal is cleaner with less channel mismatch. To illustrate the result, two segments of flaw signal are plotted.

The baseline estimation algorithm proposed in [13] is also compared in Tables 1 and 2. From Table 1, it can be seen that the flaw in show is typical small size flaw. The PSNR results in Table 2 also indicate that our algorithm minimizes channel-to-channel mismatches.

From Figures 7 and 10, it is clear that the signal has great channel mismatches which makes it impossible to evaluate the flaw size. Using median algorithm corrected data shown as Figures 8 and 11, the flaw signal is prominent and the signal of normal channel is smooth, but still some ripple exists.

Table 1: SD of raw data and corrected data.

Index of flaw	SD of raw data	SD of median corrected data	SD of baseline estimation equalized data	SD of adaptive channel equalized data	flaw size (length ∗ width ∗ depth, unit mm)
1	0.0544	0.0129	0.0121	0.0093	$40 \times 20 \times 1$
2	0.0548	0.0138	0.0136	0.0108	$\phi 6$ perforation
3	0.0565	0.0175	0.0169	0.0158	$\phi 8$ perforation
4	0.0617	0.0324	0.0318	0.0288	$20 \times 40 \times 4$
5	0.0564	0.0182	0.0179	0.0160	$40 \times 20 \times 4$
6	0.0560	0.0143	0.0143	0.0142	$20 \times 20 \times 2$
7	0.0544	0.0115	0.0104	0.0089	$40 \times 20 \times 2$
8	0.0573	0.0174	0.0163	0.0141	$20 \times 40 \times 2$

Table 2: PSNR of raw data and corrected data.

Index of flaw	PSNR of raw data	PSNR of median corrected data	PSNR of baseline estimation equalized data	PSNR of adaptive channel equalized data	flaw size (length ∗ width ∗ depth, unit mm)
1	16.8393	25.6999	26.0056	26.7182	$40 \times 20 \times 1$
2	18.6216	30.4281	31.0981	32.5947	$\phi 6$ perforation
3	19.0502	29.0965	29.2674	30.0454	$\phi 8$ perforation
4	18.5951	24.0668	24.1328	24.6181	$20 \times 40 \times 4$
5	17.5086	25.5968	25.8254	26.5034	$40 \times 20 \times 4$
6	16.8988	21.7448	22.9136	24.0683	$20 \times 20 \times 2$
7	16.4687	26.5349	26.8903	28.1939	$40 \times 20 \times 2$
8	17.5993	26.5228	26.8162	27.8077	$20 \times 40 \times 2$

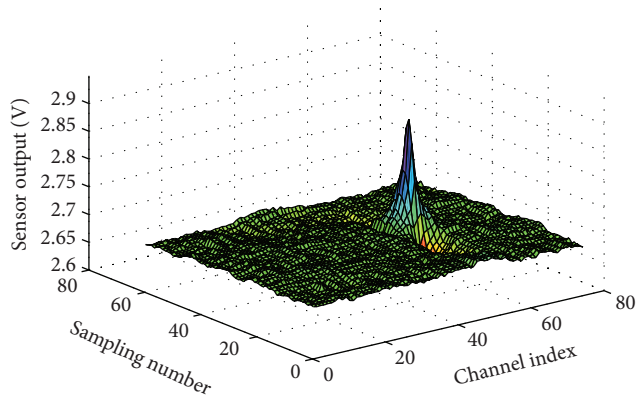


Figure 12: PCA and ELM based algorithm treated data of flaw 2.

Figures 9 and 12 show the results of algorithm proposed in this paper, the flaw signal is more prominent and the normal part is smoother, which results easy to evaluate the flaw. The SD shown in Table 1 also indicates that data treated by our algorithm has less channel mismatches and noise.

4. Conclusion

In this paper, a new adaptive channel equalization is proposed for processing of MFL signal prior to flaw characterization. The scheme performs channel equalization by using single layer neural networks, and the fast learning algorithm of ELM is used to give excellent processing speed. Focusing on the signal of flaw, a PCA based flaw detect algorithm is given to locate the flaw signal. The algorithm proposed in this paper minimizes the channel-to-channel mismatch and reduces the distortion of the signal of the flaw. Both theory analysis and simulation results show the efficiency of our algorithm. For the flaw signal, the algorithm proposed in this paper needs to locate the flaw first. For shallow and small flaw, how to locate it and make less false detection is still a problem in both theory and engineering.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61034005, 61104021), the National High Technology Research and Development Program of China (2012AA040104), the Fundamental Research Funds for the Central Universities of China (N120404022), and the Natural Science Foundation of Liaoning province, China (2013020043).

References

[1] A. A. Carvalho, J. M. A. Rebello, L. V. S. Sagrilo, C. S. Camerini, and I. V. J. Miranda, "MFL signals and artificial neural networks applied to detection and classification of pipe weld defects," *NDT & E International*, vol. 39, no. 8, pp. 661–667, 2006.

[2] R. Christen and A. Bergamini, "Automatic flaw detection in NDE signals using a panel of neural networks," *NDT and E International*, vol. 39, no. 7, pp. 547–553, 2006.

[3] Y. Xiang and S. K. Tso, "Detection and classification of flaws in concrete structure using bispectra and neural networks," *NDT and E International*, vol. 35, no. 1, pp. 19–27, 2002.

[4] S. Mukhopadhyay and G. P. Srivastava, "Characterization of metal loss defects from magnetic flux leakage signals with discrete wavelet transform," *NDT and E International*, vol. 33, no. 1, pp. 57–65, 2000.

[5] D. Mukherjee, S. Saha, and S. Mukhopadhyay, "Inverse mapping of magnetic flux leakage signal for defect characterization," *NDT & E International*, vol. 54, pp. 198–208, 2013.

[6] S. Kathirmani, A. K. Tangirala, S. Saha, and S. Mukhopadhyay, "Online data compression of MFL signals for pipeline inspection," *NDT and E International*, vol. 50, pp. 1–9, 2012.

[7] Q. He, R. Yan, F. Kong, and R. Du, "Machine condition monitoring using principal component representations," *Mechanical Systems and Signal Processing*, vol. 23, no. 2, pp. 446–466, 2009.

[8] R. Yan, R. X. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: a review with applications," *Signal Processing*, vol. 96, pp. 1–15, 2014.

[9] B. Li and X. Chen, "Wavelet-based numerical analysis: a review and classification," *Finite Elements in Analysis and Design*, vol. 81, pp. 14–31, 2014.

[10] Z. K. Zhu, R. Yan, L. Luo, Z. H. Feng, and F. R. Kong, "Detection of signal transients based on wavelet and statistics for machine fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 23, no. 4, pp. 1076–1097, 2009.

[11] R. Yan and R. X. Gao, "An efficient approach to machine health diagnosis based on harmonic wavelet packet transform," *Robotics and Computer-Integrated Manufacturing*, vol. 21, no. 4-5, pp. 291–301, 2005.

[12] Y. Zhang, Z. Ye, and X. Xu, "An adaptive method for channel equalization in MFL inspection," *NDT & E International*, vol. 40, no. 2, pp. 127–139, 2007.

[13] D. Mukherjee, S. Saha, and S. Mukhopadhyay, "An adaptive channel equalization algorithm for MFL signal," *NDT and E International*, vol. 45, no. 1, pp. 111–119, 2012.

[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[15] D. Lowe, "Adaptive radial basis function nonlinearities, and the problem of generalisation," in *Proceeding of the 1st IEE International Conference on Artificial Neural Networks*, pp. 171–175, London, UK, October 1989.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[17] G. Huang, L. Chen, and C. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 879–892, 2006.

[18] G. Li, C. F. Alcala, S. J. Qin, and D. Zhou, "Generalized reconstruction-based contributions for output-relevant fault diagnosis with application to the Tennessee Eastman process," *IEEE Transactions on Control Systems Technology*, vol. 19, no. 5, pp. 1114–1127, 2011.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 198362, 15 pages
http://dx.doi.org/10.1155/2014/198362

Research Article

Time-Frequency Fault Feature Extraction for Rolling Bearing Based on the Tensor Manifold Method

Fengtao Wang,[1] Shouhai Chen,[1] Jian Sun,[1] Dawen Yan,[2] Lei Wang,[1] and Lihua Zhang[3]

[1] *Institute of Vibration Engineering, School of Mechanical Engineering, Dalian University of Technology, Dalian 116024, China*
[2] *School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China*
[3] *Institute of Microelectromechanical Systems and Precision Engineering, School of Mechanical Engineering, Dalian University of Technology, Dalian 116024, China*

Correspondence should be addressed to Fengtao Wang; wangft@dlut.edu.cn

Received 2 May 2014; Revised 20 June 2014; Accepted 10 July 2014; Published 4 August 2014

Academic Editor: Weihua Li

Rolling-bearing faults can be effectively reflected using time-frequency characteristics. However, there are inevitable interference and redundancy components in the conventional time-frequency characteristics. Therefore, it is critical to extract the sensitive parameters that reflect the rolling-bearing state from the time-frequency characteristics to accurately classify rolling-bearing faults. Thus, a new tensor manifold method is proposed. First, we apply the Hilbert-Huang transform (HHT) to rolling-bearing vibration signals to obtain the HHT time-frequency spectrum, which can be transformed into the HHT time-frequency energy histogram. Then, the tensor manifold time-frequency energy histogram is extracted from the traditional HHT time-frequency spectrum using the tensor manifold method. Five time-frequency characteristic parameters are defined to quantitatively depict the failure characteristics. Finally, the tensor manifold time-frequency characteristic parameters and probabilistic neural network (PNN) are combined to effectively classify the rolling-bearing failure samples. Engineering data are used to validate the proposed method. Compared with traditional HHT time-frequency characteristic parameters, the information redundancy of the time-frequency characteristics is greatly reduced using the tensor manifold time-frequency characteristic parameters and different rolling-bearing fault states are more effectively distinguished when combined with the PNN.

1. Introduction

Rolling bearings are widely used in modern rotating machinery, and their failure is one of the most common causes of machine breakdowns and accidents [1–3]. Therefore, fault diagnosis of rolling bearings is necessary to ensure the safe and efficient operation of machines in engineering applications. The main aspects of bearing fault diagnosis are classification and pattern recognition, where feature extraction directly affects the accuracy and reliability of the fault diagnosis [4]. Rolling-bearing fault features can be generally divided into three categories: time-domain characteristics, frequency-domain characteristics, and time- and frequency-domain characteristics [5, 6].

Time-domain characteristics are fairly intuitive; however, they fluctuate significantly and lack quantitative judging criteria. Thus, they cannot be directly used to diagnose bearing faults. In contrast, frequency-domain characteristics can be used to diagnose bearing fault conditions more accurately because different bearing faults correspond to different characteristic frequencies. However, there are typically noise and modulation components in the bearing fault signals. Thus, direct application of the frequency-domain method will submerge the fault characteristic frequency in noise or false frequency components because of improper selection of the demodulation parameters. Furthermore, signal denoising and demodulation must be conducted before extracting the bearing fault characteristic frequencies. In the process of signal denoising and demodulation, parameters such as the denoising parameters, demodulation center, and filter bandwidth should be properly selected based on experience, and a satisfactory selection is only obtained after numerous adjustments.

The time- and frequency-domain characteristics, which have the intuitive feature of the time-domain characteristics

and good time-frequency aggregation, can simultaneously reflect the time-domain and frequency-domain characteristics of a signal [7–10]. Therefore, extracting the time-frequency fault characteristics is important for fault diagnosis. Wang and Hu used the principle of time-frequency image analysis to diagnose gearbox faults in 1993 [11]; this effort was the first application of time-frequency image for the fault diagnosis of machinery and equipment. Zhang et al. subsequently used time-frequency images to classify diesel engine faults under complex vibration conditions [12]. Zhu et al. used short-time Fourier transform to extract time-frequency features for fault diagnosis [13], and satisfactory results were achieved. However, the aforementioned time-frequency characteristics are not adaptive and can only be used for reciprocating machinery. To overcome the limitations of the above methods, Huang et al. proposed the HHT time-frequency spectrum, which is self-adaptive [14]. The HHT time-frequency spectrum is suitable for analyzing nonstationary signals because of its frequency instantaneity [15]. However, mode mixing is inevitable for signals with instantaneous frequency trajectory crossings [16, 17]. Li et al. used the geometric center of the HHT time-frequency spectrum as a feature vector [18, 19] in combination with SVM and classified rolling-bearing fault signals. However, because the geometric center requires a considerable amount of calculations and lacks corresponding physical meaning, it can only provide qualitative classification criteria. Manifold learning has recently emerged in nonlinear-feature extraction because of its capability of effectively identifying hidden low-dimensional nonlinear structures in high-dimensional data. He [20] proposed a time-frequency manifold feature by combining the time-frequency distribution and the nonlinear manifold for an effective quantitative representation of machinery health pattern.

This paper proposes a new tensor manifold time-frequency feature extraction method to overcome the weakness of traditional HHT time-frequency characteristics. The HHT time-frequency spectrum, which contains a considerable amount of failure information, is used as the research object. The tensor manifold learning method is applied to extract the tensor manifold time-frequency characteristics of the HHT time-frequency spectrum. The two-dimensional time-frequency information does not need to be converted into a one-dimensional vector when calculating the tensor manifold, and the information loss is significantly reduced. On this basis, five time-frequency characteristic parameters are defined. The tensor manifold time-frequency characteristic parameters can distinguish different rolling-bearing fault states more effectively than traditional HHT time-frequency characteristic parameters. Combined with PNN, the tensor manifold time-frequency characteristic parameters can effectively distinguish different rolling-bearing fault states. Engineering vibration signals were used to evaluate the efficiency of the proposed method.

The remainder of this paper is organized as follows. The theory basis is introduced in Section 2, and the tensor manifold time-frequency fault feature extraction method is described in Section 3. Section 4 presents the adaption of the proposed method to rolling-bearing fault classification. Rolling-bearing fault classification is implemented in Section 5. Finally, conclusions are drawn in Section 6.

2. Theory Basis

2.1. HHT Time-Frequency Spectrum. Based on the definition of instantaneous frequency and EMD, the HHT time-frequency spectrum is analytically derived as follows.

Apply the EMD to signal $X(t)$ to obtain the IMFs of $X(t)$. Then, the analytical form of $X(t)$ can be expressed as

$$X(t) = \mathrm{Re}\sum_{i=1}^{n} A_i(t)\, e^{j\int \omega_i(t)dt}, \tag{1}$$

where Re is the real part of the selected signal, $A_i(t)$ is the instantaneous amplitude of the ith IMF, and $\omega_i(t)$ is the corresponding instantaneous frequency.

The time, frequency, and amplitude of the signals can be combined to form the three-dimensional time-frequency space. Then, the amplitude distribution on time-frequency plane is referred to as the HHT time-frequency spectrum, which is expressed as

$$H(t,\omega) = \mathrm{Re}\sum_{i=1}^{n} b_i A_i(t)\, e^{j\int \omega_i(t)dt}, \tag{2}$$

where Re is the real part of the selected signal and b_i is the indicator variable. When $\omega_i = \omega$, $b_i = 1$, and when $\omega_i \neq \omega$, $b_i = 0$.

The HHT time-frequency analysis is a decomposition method based on signal local characteristics, which provides a physical basis for the concept of instantaneous frequency and sets this method apart from conventional methods through its use of numerous harmonic components to describe complex nonlinear and nonstationary signals. Therefore, from the concept definition and the nature of signal analysis, the HHT time-frequency spectrum eliminates the limitations of Fourier transform and can accurately describe nonstationary signal characteristics.

2.2. Tensor Manifold Algorithm

2.2.1. Locality Preserving Projection (LPP) Manifold Learning Algorithm. The LPP manifold learning algorithm aims at finding the linear transformation matrix $\mathbf{W}$ to reduce the dimensionality of high-dimensional data. There are l training samples $\{\mathbf{x}_i\}_{i=1}^{l} \in \mathbf{R}^m$, and $\mathbf{W}$ can be obtained by minimizing the following objective function:

$$\min_{\mathbf{W}} \left(\sum_{i,j} \left(\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right)^2 S_{i,j} \right), \tag{3}$$

where $S_{i,j}$ is the similarity measure among objects and can be defined using the k-nearest-neighbor method:

$$S_{i,j} = \begin{cases} \exp\left(-\dfrac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_f^2}{t}\right), & \text{if } \mathbf{x}_i \in O(k,\mathbf{x}_i) \text{ or } \mathbf{x}_j \in O\left(k,\mathbf{x}_j\right) \\ 0, & \text{Otherwise} \end{cases}, \tag{4}$$

where $O(k, \mathbf{x}_i)$ denotes the k nearest neighbor of $\mathbf{x}_i$ and t is a positive constant. Both k and t can be determined empirically.

Equation (3) demonstrates the feature space after dimension reduction can maintain the local structure of the original high-dimensional space. We apply an algebraic transformation to (3) as follows:

$$\begin{aligned}&\frac{1}{2}\sum_{i,j}\left(\mathbf{W}^T\mathbf{x}_i-\mathbf{W}^T\mathbf{x}_j\right)^2 S_{ij}\\&=\sum_{i,j}\mathbf{W}^T\mathbf{x}_i D_{ii}\mathbf{x}_i^T\mathbf{W}-\sum_{i,j}\mathbf{W}^T\mathbf{x}_i S_{ij}\mathbf{x}_i^T\mathbf{W}\\&=\mathbf{W}^T\mathbf{X}\left(\mathbf{D}-\mathbf{S}\right)\mathbf{X}^T\mathbf{W}=\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W},\end{aligned}\tag{5}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l]$, $\mathbf{D}$ denotes an $l \times l$ diagonal matrix, where the diagonal element $D_{ii} = \sum_i S_{ij}$, $\mathbf{S} = (S_{ij})_{l\times l}$, and $\mathbf{L} = \mathbf{D} - \mathbf{S}$.

Then, the problem of solving for the optimal vector $\mathbf{W}$ can be transformed into the following eigenvalue problem:

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}=\lambda\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W}.\tag{6}$$

2.2.2. Tensor LPP Manifold Learning Algorithm. The LPP manifold learning algorithm [21] can only be regarded as a one-dimensional manifold feature extraction algorithm. However, the number of training images in the two-dimensional (e.g., time-frequency spectrum) image feature extraction process is notably small compared to the dimensions of the image vectors, which results in a singularity of $\mathbf{X}\mathbf{D}\mathbf{X}^T$ and failure of the LPP algorithm. To alleviate the drawback of the LPP, this paper uses a new tensor LPP manifold learning algorithm (Ten-LoPP) [22] to extract the time-frequency spectrum fault characteristics.

There are l two-dimensional training images $\{\mathbf{A}_i\}_{i=1}^{l} \in \mathbf{R}^{m\times n}$, where $\boldsymbol{\omega}$ denotes an n-dimensional unitization column vector. The main objective of the tensor manifold algorithm is to make each $m \times n$ image matrix $\mathbf{A}_i$ project onto $\boldsymbol{\omega}$ using a linear transformation $\mathbf{y}_i = \mathbf{A}_i\boldsymbol{\omega}$. In this manner, an m-dimensional column vector can be obtained and considered a projection feature vector of image $\mathbf{A}_i$. The objective function of the tensor manifold algorithm is expressed as follows:

$$\begin{aligned}&\frac{1}{2}\sum_{i,j}\left(\mathbf{A}_i\boldsymbol{\omega}-\mathbf{A}_j\boldsymbol{\omega}\right)^2 S_{ij}\\&=\sum_{i,j}\mathbf{W}^T\mathbf{A}_i^T D_{ii}\mathbf{I}_m\mathbf{A}_i\boldsymbol{\omega}-\sum_{i,j}\mathbf{W}^T\mathbf{A}_i^T S_{ij}\mathbf{I}_m\mathbf{A}_i\boldsymbol{\omega}\\&=\mathbf{W}^T\mathbf{A}^T\left[\left(\mathbf{D}-\mathbf{S}\right)\otimes\mathbf{I}_m\right]\mathbf{A}\mathbf{W}=\mathbf{W}^T\mathbf{A}^T\left(\mathbf{L}\otimes\mathbf{I}_m\right)\mathbf{A}\mathbf{W},\end{aligned}\tag{7}$$

where $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_l]$, the definitions of $\mathbf{D}$ and $\mathbf{L}$ are identical to those in the LPP manifold learning method, and $\otimes$ denotes the Kronecker product.

Then, the problem of solving for the optimal vector $\boldsymbol{\omega}$ is transformed into the following eigenvalue problem:

$$\mathbf{A}^T\left(\mathbf{L}\otimes\mathbf{I}_m\right)\mathbf{A}\boldsymbol{\omega}=\lambda\mathbf{A}^T\left(\mathbf{D}\otimes\mathbf{I}_m\right)\mathbf{A}\boldsymbol{\omega},\tag{8}$$

where $\boldsymbol{\omega}$ is comprised of d feature vectors that correspond to the smallest nonzero eigenvalues; that is, there are d optimal projection vectors $\boldsymbol{\omega}$, which can form the projection matrix $\mathbf{W} = [\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_d]$. For any image $\mathbf{A}_x$, there is

$$\mathbf{y}_{xi}=\mathbf{A}_x\boldsymbol{\omega}_i,\quad i=1,2,\ldots,d,\tag{9}$$

where $\mathbf{y}_{x1}, \mathbf{y}_{x2}, \ldots, \mathbf{y}_{xd}$ are the projection feature vectors of the sample image $\mathbf{A}_x$ and $\mathbf{y}_x = [\mathbf{y}_{x1}, \mathbf{y}_{x2}, \ldots, \mathbf{y}_{xd}]$, which is comprised of projection feature vectors, is the characteristic matrix of the sample image $\mathbf{A}_x$.

2.3. Probabilistic Neural Network (PNN). The neural composition structure and elements of the PNN are shown in Figure 1

In the PNN, characteristic parameters were transported into each node on the pattern layer through the input layer. Then, we apply layer nonlinear mapping to the input parameters in each node of the PNN pattern and complete the comparison between an unknown type with a known type. Finally, the characteristic parameters that represent the types are input to the next layer for processing. The node structure of the layers is used to be called the RBF center, and the node output is expressed as follows:

$$\mathbf{O}_i=\mathbf{R}_i\left(\left\|\mathbf{X}-\boldsymbol{\omega}_i\right\|\right),\tag{10}$$

where the ith center vector is $\boldsymbol{\omega}_i$, which is the same size as the input vector. $R_i(\cdot)$ denotes the radial basis function, which is typically a Gaussian function; that is,

$$\exp\left(-\frac{\left\|\mathbf{X}-\boldsymbol{\omega}_i\right\|^2}{2\sigma_i^2}\right),\tag{11}$$

where σ_i denotes the shape parameter that corresponds to the ith component of the radial basis function.

To facilitate the calculation, $\mathbf{X}$ and $\boldsymbol{\omega}_i$ are processed with mathematical regularization and unit. Assume that $\mathbf{z}_i = \mathbf{X}\cdot\boldsymbol{\omega}_i$. Then, the above expression is expressed as follows:

$$g\left(\mathbf{z}_i\right)=\exp\left[\frac{\left(\mathbf{z}_i-1\right)}{\sigma^2}\right].\tag{12}$$

Finally, through the output layer (decision-making layer), the characteristic parameters, which are derived from the pattern layer, are accumulated to provide the category feature vector, which is

$$f_A\left(\mathbf{X}\right)=\sum_{j=1}^{N}g\left(\mathbf{z}_i\right).\tag{13}$$

The PNN has the following characteristics: (1) the training convergence speed is high, making the PNN suitable for the real-time processing of various data types; (2) the pattern unit can form any nonlinear mapping judgment surface, which is closest to the optimal judgment surface bayes; (3) the selection of the RBF center kernel function has diversity, and the form of the kernel function has a small effect on the recognition results; and (4) the number of neuron nodes in each PNN layer is relatively stable, the hardware processing is convenient, and the fault tolerance is high. The PNN has been widely used in pattern recognition, prediction estimation, and filtering denoising.

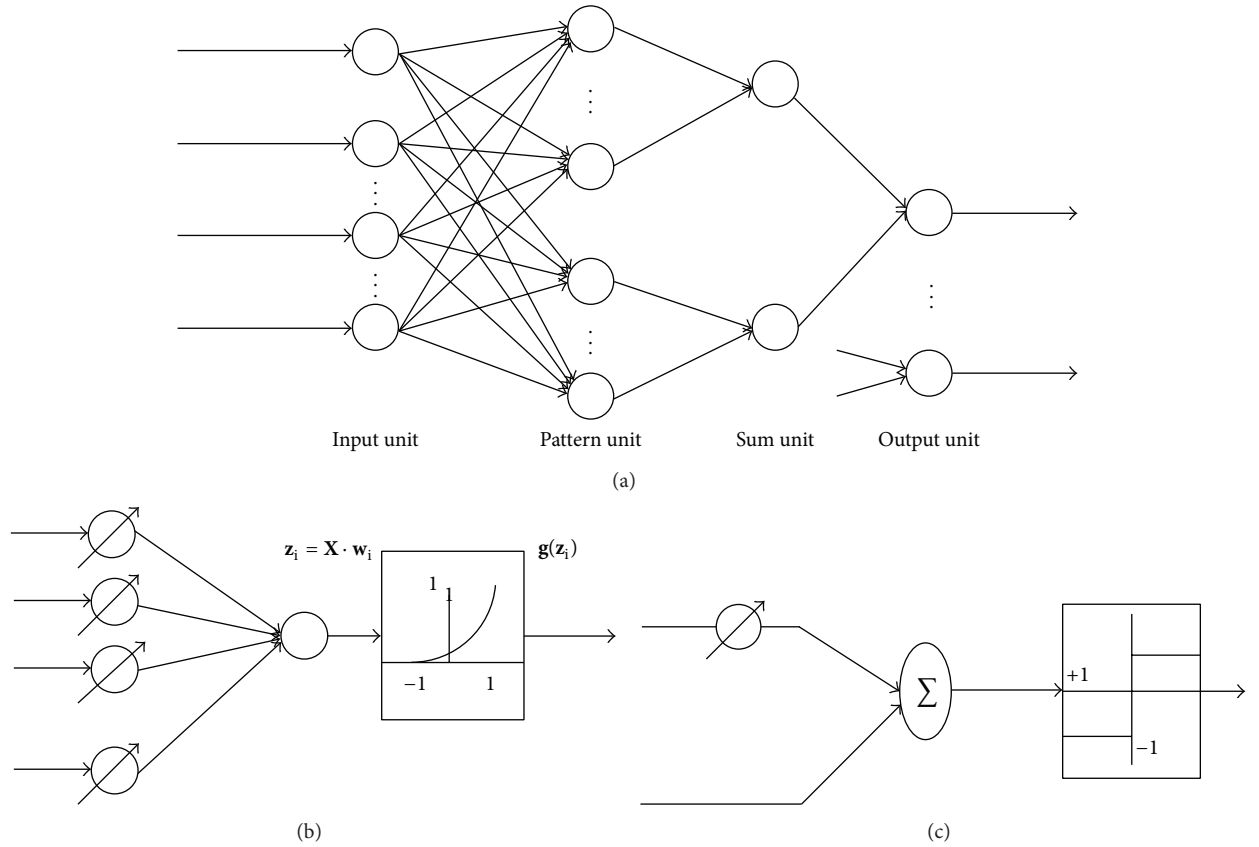


Figure 1: Internal composition structure and elements of the PNN: (a) system structure of the PNN, (b) pattern unit of the PNN, and (c) output unit of the PNN.

3. Tensor Manifold Time-Frequency Fault Feature Extraction Method

3.1. Description of the Proposed Method. The manifold learning method is a nonlinear dimension reduction method to extract low-dimensional nonlinear characteristics from high-dimensional data. Unlike the conventional linear dimension reduction methods, such as multidimensional scaling (MDS), principal component analysis (PCA), and linear discriminant analysis (LDA), this method is a nonlinear method to address the part before the whole. By satisfying the entire optimization, the manifold learning method can preserve the partial manifold characteristics and effectively extract the nonlinear manifold characteristics that are inherent in the high-dimensional characteristic set. However, the manifold learning algorithm suffers from information loss and error that are caused by the transformation from a set of two-dimensional time-frequency characteristics to a one-dimensional vector.

To alleviate the drawback of information loss and error, this section presents a tensor manifold time-frequency fault feature extraction method based on the tensor manifold algorithm to extract the set of low-dimensional time-frequency characteristics from the set of high-dimensional time-frequency characteristics. Then, five tensor manifold time-frequency characteristic parameters were defined and combined with the PNN to classify the rolling-bearing failure samples.

The tensor manifold time-frequency fault feature extraction method is described as follows, and Figure 2 presents its flow chart.

(1) Group the rolling-bearing vibration signal samples to be classified and for training and then calculate the HHT time-frequency spectrum. To hasten the calculation of the tensor manifold algorithm, grid the time-frequency regions, integrate the energy value of the HHT time-frequency spectrum of each mesh, and convert the HHT time-frequency spectrum into HHT time-frequency energy histograms.

(2) The HHT time-frequency energy histograms are essentially two-dimensional matrices. Use the HHT time-frequency energy histograms that correspond to signal samples to form a set of high-dimensional time-frequency characteristics.

(3) Apply the tensor manifold algorithm to extract the set of low-dimensional time-frequency characteristics from the set of high-dimensional time-frequency characteristics. In this manner, the tensor manifold time-frequency energy histograms are obtained.

(4) Based on the result of step (3), define different tensor manifold time-frequency characteristic parameters.

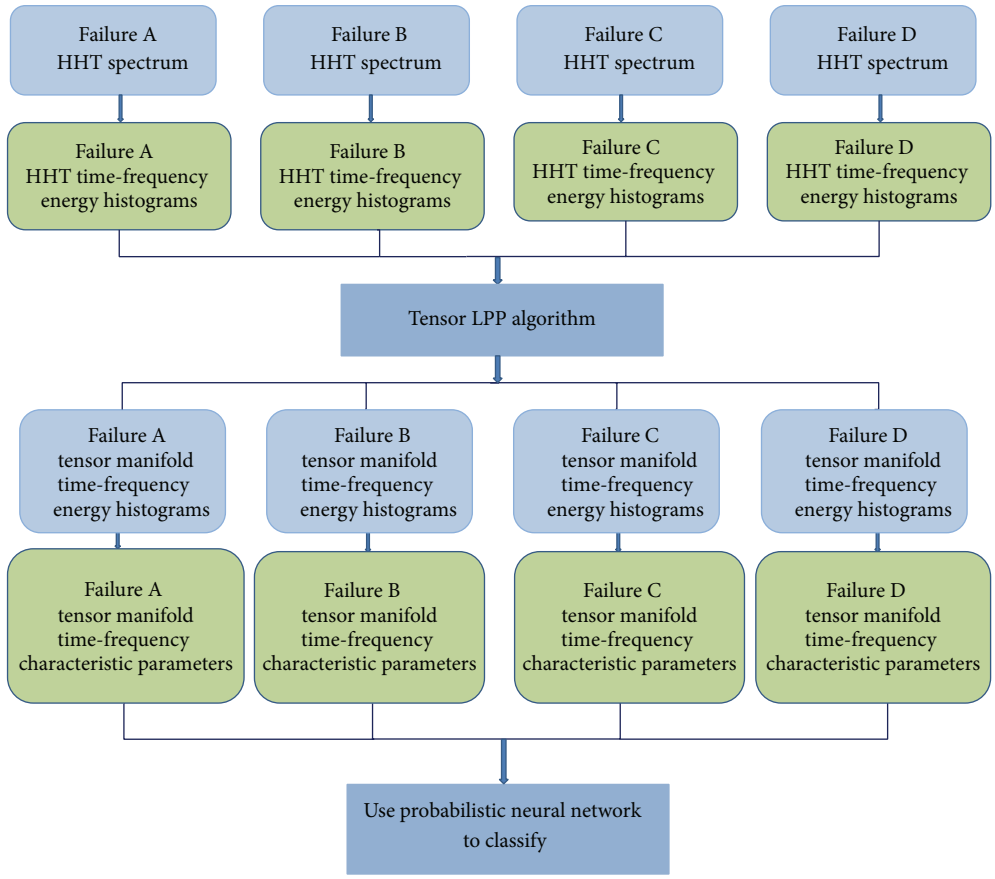


Figure 2: Flow chart of the time-frequency characteristic extraction method based on a tensor manifold.

Input the defined parameters of the training signal samples into the PNN for the rolling-bearing fault classification.

(5) Input the tensor manifold time-frequency characteristic parameters of the to-be-classified signal samples into the trained PNN to classify the rolling-bearing faults.

3.2. Definition of the Time-Frequency Characteristic Parameters. The tensor manifold time-frequency energy histogram is a nonlinear time-frequency fault feature and can effectively differentiate different rolling-bearing fault signals. However, it is equal to a two-dimensional matrix, which makes it unsuitable for direct application in fault classification. In this section, several parameters are presented to quantitatively measure the difference among the tensor manifold time-frequency energy histograms. Their definitions are provided as follows.

3.2.1. Energy Entropy. Entropy is proposed to measure the data complexity and the probability to generate the new signal model. Here, the energy entropy is defined as follows:

$$H = -\sum_{i=1}^{n} p_i \log\left(p_i\right), \qquad p_i = \frac{e_i}{\sum_{i=1}^{n} e_i}, \tag{14}$$

where H is the energy entropy, e_i is the value of the time frequency energy histogram, and p_i is the proportion of each e_i in the total $\sum_{i=1}^{n} e_i$. In addition, the energy entropy can reflect the uncertainty in the energy distribution.

3.2.2. Energy Correlation Coefficient. Divide the time-frequency energy histogram into m sections by frequency and mark $\mathbf{E}_{f1}, \mathbf{E}_{f2}, \ldots, \mathbf{E}_{fm}$ and $\mathbf{E}_t = \mathbf{E}_{f1} + \mathbf{E}_{f2} + \cdots + \mathbf{E}_{fm}$. Because each $\mathbf{E}_{fi}$ varies with different time-frequency energy histograms, we can analyze the relevance of $\mathbf{E}_{fi}$ and $\mathbf{E}_t$ to measure the difference in the time-frequency energy histogram. The energy correlation coefficient vector is defined as follows:

$$\mathbf{Ecoef} = \left[\text{corcoef}(1),\ \text{corcoef}(2),\ \cdots\ \text{corcoef}(m)\right]^T, \tag{15}$$

where $\text{corcoef}(i) = \text{corcoef}(\mathbf{E}_{fi}, \mathbf{E}_t)$ and $\text{corcoef}(\cdot)$ is the cross-correlation function.

3.2.3. Energy Sparsity. The signal energy distribution of the time-frequency energy histogram varies more significantly as it approaches zero. The sparsity expresses the sparse distribution of energy, and the purpose of estimating the sparsity is to obtain a function $q(\mathbf{x})$, $\mathbf{x} \in \mathbf{R}^n$. If $\mathbf{x}$ is sparse, then $q(\mathbf{x})$ is relatively large, and vice versa. Generally, the norm L_p of vector $\mathbf{x}$ is used to quantitatively estimate the sparsity. Here,

we define the L_p norm of the standardized form of vector $\mathbf{x}$ as follows:

$$q^{-1}(\mathbf{x}) = \frac{\|\mathbf{x}\|_p}{n^{1/p-1/2} \cdot \|\mathbf{x}\|_2} = \frac{1}{n^{1/p-1/2}} \cdot \frac{\left(\sum_{k=1}^{n} \mathbf{x}_k^p\right)^{1/p}}{\left(\sum_{k=1}^{n} \mathbf{x}_k^2\right)^{1/2}}, \quad (16)$$

where $1 \leq p < \infty$, and we select $p = 1$ such that L_1 can accurately reflect the energy distribution of the histogram.

3.2.4. Energy Mutual Information. Mutual information is proposed to measure the degree of independence among random variables. The mutual information of multiple variables is defined as the KL divergence of the multivariate joint probability density and its marginal probability density product:

$$\begin{aligned} I(x) &= KL\left(p(x), \prod_{i=1}^{N} p_i(x_i)\right) \\ &= \int p(x) \log\left(\frac{p(x)}{\prod_{i=1}^{N} p_i(x_i)}\right) dx, \end{aligned} \quad (17)$$

where KL is the divergence, $x = x_1, x_2, \ldots x_N$, $p(x)$ is the multivariate joint probability density function, and $p_i(x_i)$, $(i = 1 \sim N)$ is the marginal probability density function. Then KL is defined as follows:

$$KL[p(x), q(x)] = \int p(x) \log \frac{p(x)}{q(x)} dx, \quad (18)$$

where $p(x)$, $q(x)$ are two different probability density functions of a random vector x. The energy mutual information can be calculated for each histogram according to (18).

3.2.5. Energy Kurtosis. Kurtosis is a physical parameter that is proposed to measure the degree of Gaussian distribution of a random variable. A larger energy kurtosis in the time-frequency energy histogram corresponds to weaker Gaussianity of the energy distribution, whereas a smaller kurtosis indicates stronger Gaussianity. If the Gaussianity of the energy distribution is strong, the energy distribution presents the "middle big, two sides small" phenomenon. The energy values are mainly within the middle range, and larger or smaller values are less likely to occur. For the sequence $\mathbf{X} = x_1, x_2, \ldots x_N$ of energy values, the overall kurtosis is defined as

$$\text{Kurt}(x_{1,2,\ldots,N}) = \frac{\sum_{i=1}^{N} (x_i - \mu)/N}{\sigma^4} - 3. \quad (19)$$

4. Application of the Proposed Method to Rolling-Bearing Fault Classification

4.1. Rolling-Bearing Fault Data. To verify the effectiveness of the proposed method, the new method was used to analyze bearing fault data from the Bearing Data Center of Case Western Reserve University [23].

The test rig, which is shown in Figure 3, was constructed for the run-to-failure testing of the rolling bearing. A 1.5 kW 3-phase induction motor was connected to a power meter and torque sensor by self-calibration coupling, which drove the fan. The load was adjusted using the fan. Data were collected using a vibration acceleration sensor, which was vertically fixed above the chassis of the drive end bearings of the induction motor. The bearings are deep-groove ball bearings of the type SKF6205-2RS JEM. There is a single point of failure in the inner ball and outer surface of the machining spark; the failure sizes are 0.18 mm in diameter and 0.28 mm deep. The experimental data were collected with a sample frequency of 12,000 Hz and a shaft running speed of 29.53 Hz (1,772 rpm). The corresponding ball pass frequency inner-race (BPFI), ball rotation frequency (BS), and ball pass frequency outer-race (BPFO) were estimated to be 159.93, 139.19, and 105.87 Hz, respectively.

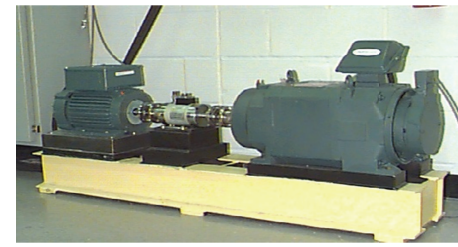


Figure 3: Diagram of the experimental test rig.

Figures 4(a)–4(d) depict a group of normal, inner-race fault, ball fault, and outer-race fault time-domain signals. Although there are several differences among the four types of signals in the time-domain wave nature, it is difficult to distinguish the rolling-bearing fault conditions using these intuitive qualitative differences. Therefore, the fault features that quantitatively represent the differences of different rolling-bearing fault statuses must be studied.

4.2. HHT Time-Frequency Characteristics of Rolling Bearings. First, the HHT time-frequency spectra of four states were calculated using the HHT method. Then, the HHT time-frequency spectrum was divided into 64 regions of identical size. The histogram of the HHT time-frequency spectrum was obtained via the integral of the energy amplitude for each region. Different types of signal HHT time-frequency spectra and their histograms are shown in Figure 5.

Figure 5 describes the HHT time-frequency spectrum and corresponding time-frequency energy histograms of the normal rolling-bearing vibration signal. As shown in Figure 5, the time-frequency energy is mainly distributed in the low-frequency region and decreases with increasing frequency. The amplitude ranges from 0 to 20 g^2.

Figure 6 presents the HHT time-frequency spectrum and corresponding time-frequency energy histogram with an inner-race fault. As shown in Figure 6, the time-frequency energy is widely distributed in both the low- and high-frequency regions, with a lull in the mid-frequency region. The amplitude ranges from 0 to 40 g^2, the maximum value of which is greater than that in the normal case.

Figure 7 presents the HHT time-frequency spectrum and corresponding time-frequency energy histograms with a ball

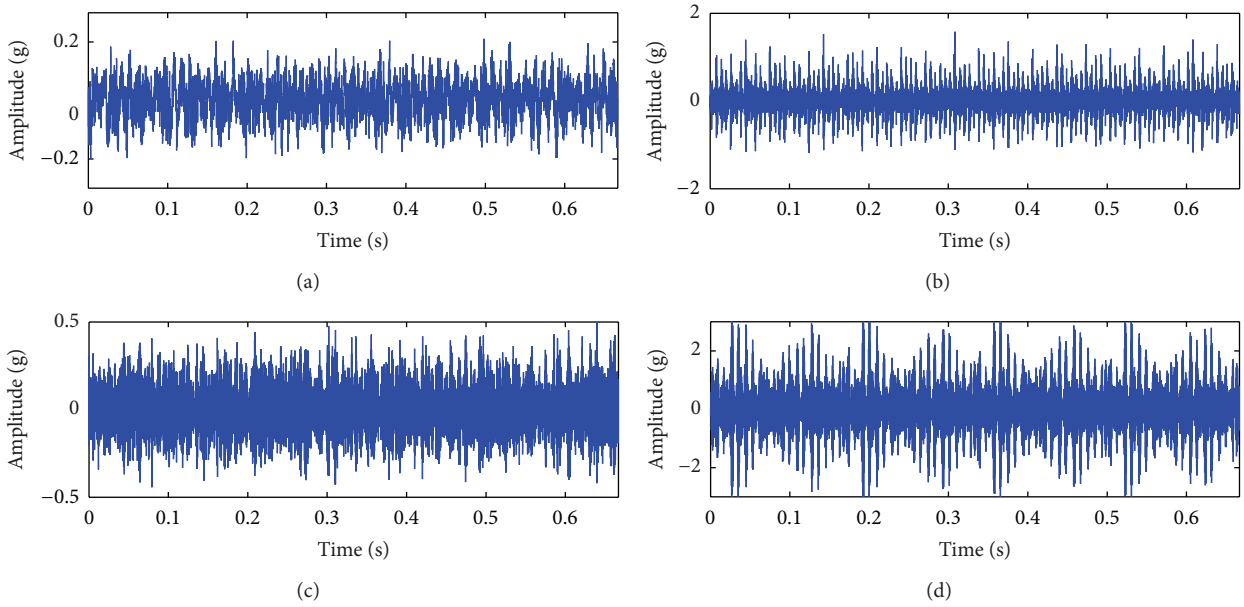


Figure 4: Four heterogeneous rolling-bearing fault data: (a) normal, (b) inner-race faults, (c) ball faults, and (d) outer-race faults.

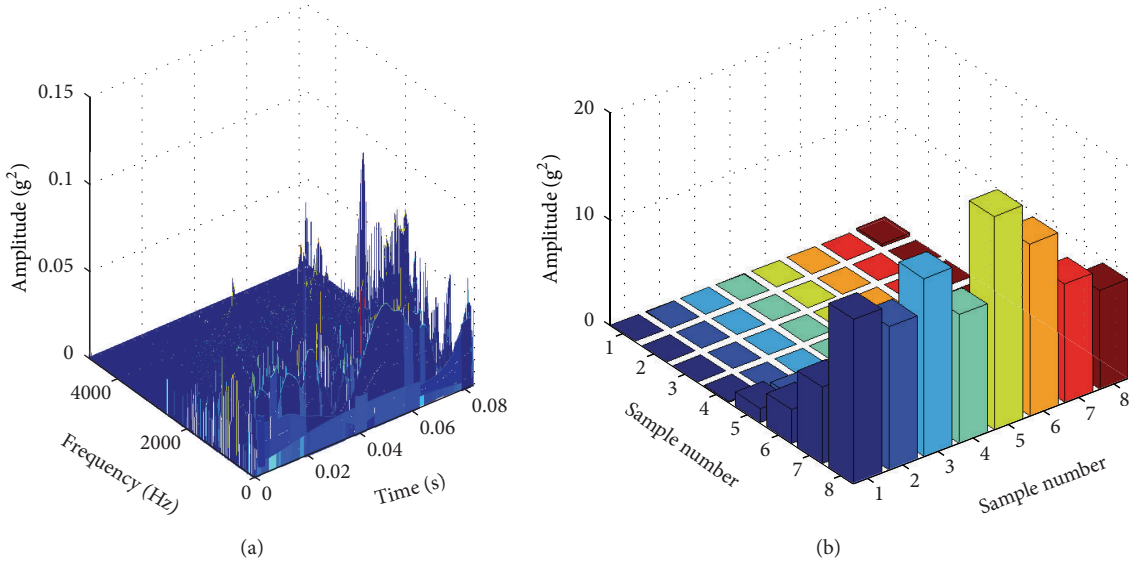


Figure 5: (a) Normal HHT time-frequency spectrum and (b) normal time-frequency energy histogram.

fault. As shown in Figure 7, the time-frequency energy is mainly distributed in the high-frequency region and exhibits a less significant yet stable distribution in the low-frequency region. The amplitude of the energy histogram ranges from 0 to 30 g^2.

Figure 8 presents the HHT time-frequency spectrum and corresponding time-frequency energy histograms with an outer-race fault. As shown in Figure 8, the time-frequency energy is centered in the high-frequency region and exhibits a lull in the low-frequency region. The distribution trend begins at a rather low frequency and increases abruptly at a certain high frequency. The magnitude ranges from 0 to 100 g^2.

4.3. Extraction of the Tensor Manifold Time-Frequency Characteristic Parameters. For convenience and conciseness, we only discuss the tensor manifold time-frequency characteristic parameters of four types of rolling-bearing faults in this section; in other situations, such as different damage degrees in the same fault type or different fault types with different damage degrees, the classification result will also be discussed in subsequent Section 5.

We consider 20 normal inner-race fault, ball fault, and outer-race fault time-domain signals (each sample dataset has 1,024 points, and 80 samples are used for training). The HHT time-frequency spectrum and HHT time-frequency energy histogram of each signal are obtained using the

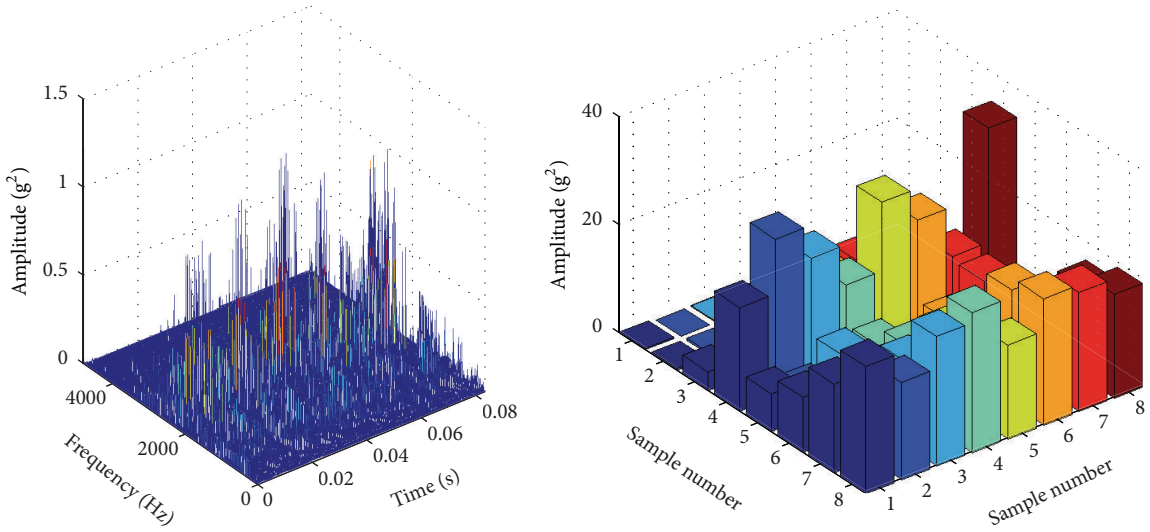


Figure 6: (a) HHT time-frequency spectrum with an inner-race fault and (b) time-frequency energy histogram with an inner-race fault.

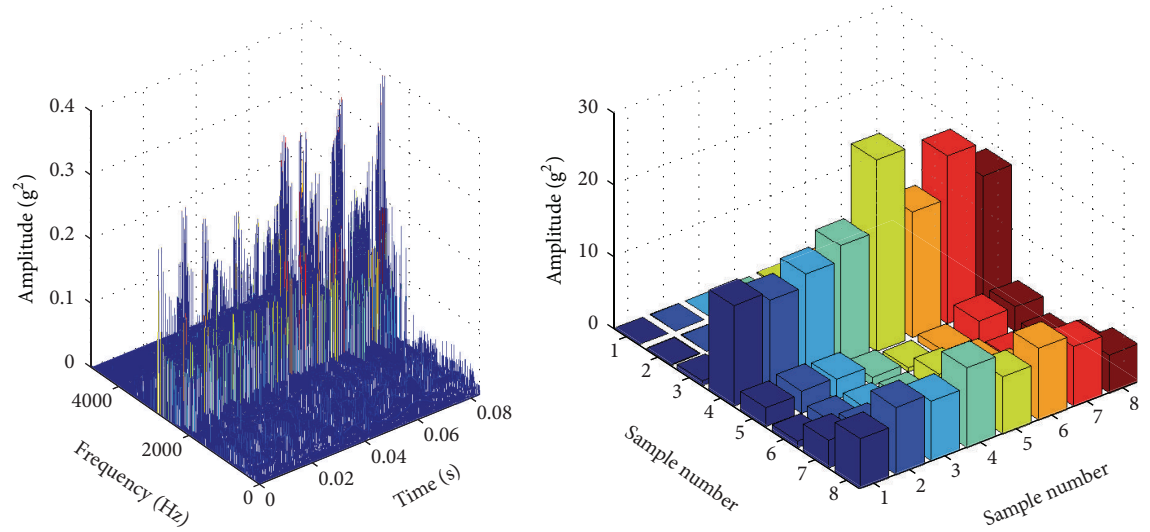


Figure 7: (a) HHT time-frequency spectrum with a ball fault and (b) time-frequency energy histogram with a ball fault.

aforementioned method. Then, a high-dimensional time-frequency feature combination with 80 samples is obtained. The tensor manifold algorithm is applied to extract the low-dimensional tensor manifold features from the high-dimensional characteristic set. The optimal projected vectors $\mathbf{W} = [\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_d]$ are obtained, and the parameter is defined as 6 because of the distribution of eigenvalues. Finally, we project the $8 * 8$ matrix energy histogram onto $\mathbf{W}$ and obtain the $8 * 6$ tensor manifold energy histograms.

According to the aforementioned definition of the five time-frequency characteristic parameters, we take the absolute value of the elements of the obtained $8*6$ tensor manifold energy histograms and calculate the tensor manifold time-frequency characteristic parameters of the tensor manifold energy histograms.

Five tensor manifold time-frequency characteristic parameters of the above 80 samples are obtained. Below, only 10 samples of the above four different types of rolling-bearing signals are considered for clarity in the graphics.

The results are as follows.

4.3.1. Manifold Energy Entropy. Samples 1–10 correspond to the normal signals, samples 11–20 correspond to the inner-race fault signals, samples 21–30 correspond to the ball fault signals, and samples 31–40 correspond to the outer-race fault signals. The manifold energy entropy of each tensor manifold time-frequency energy histogram is calculated and depicted in Figure 9(a). The energy entropy of the HHT time-frequency energy histogram without a tensor manifold analysis is presented in Figure 9(b). Compared to Figure 9(a),

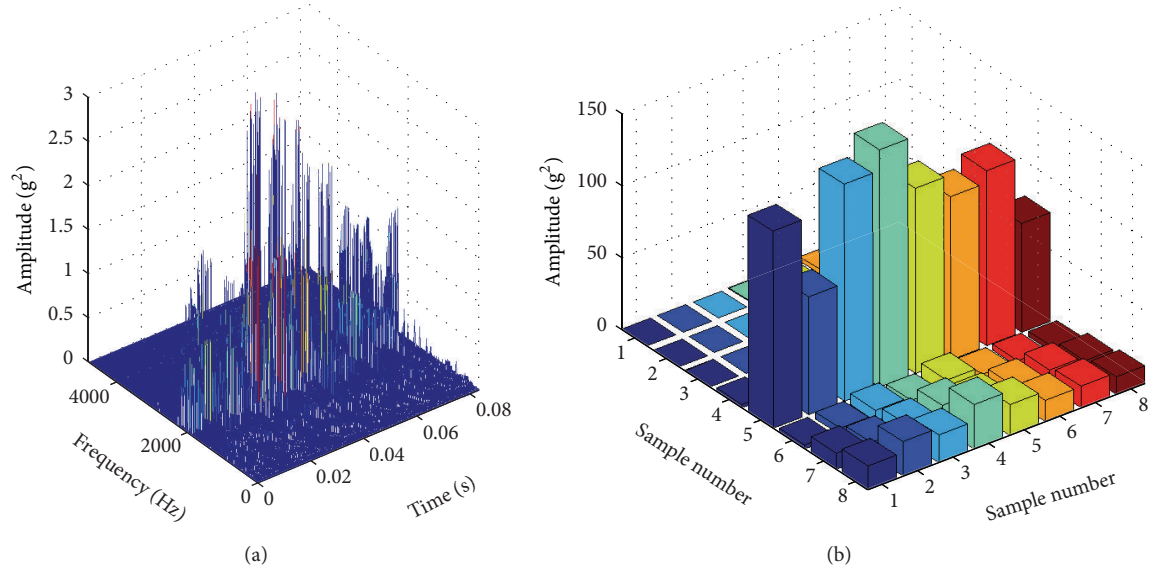


FIGURE 8: (a) HHT time-frequency spectrum with outer-race fault and (b) time-frequency energy histogram with outer-race fault.

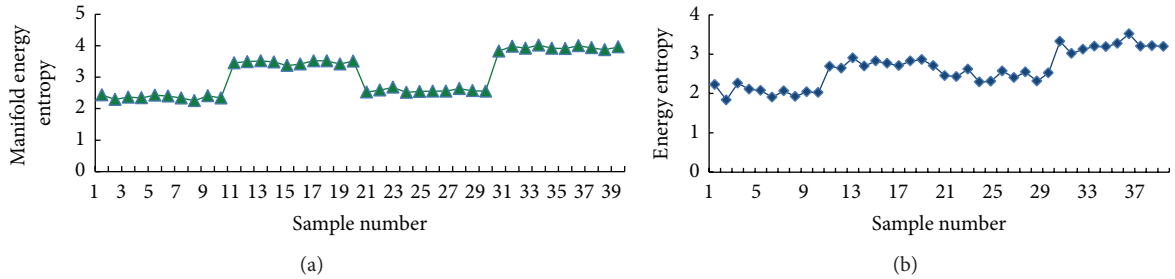


FIGURE 9: (a) Manifold energy entropy of the four fault signals and (b) energy entropy of the four fault signals.

the energy entropy shown in Figure 9(b) cannot provide a clear distinction between inner-race faults and ball faults. Thus, the manifold energy entropy is more appropriate for classifying rolling-bearing faults.

4.3.2. Manifold Energy Correlation Coefficient. The manifold energy correlation coefficient (manifold energy corcoef(i) for short) is obtained by calculating the manifold energy corcoef(i) between $\mathbf{E}_{f1}, \mathbf{E}_{f2}, \ldots, \mathbf{E}_{f6}$ and $\mathbf{E}_t$. The results are shown in Figure 10.

As shown in Figure 10, the manifold energy corcoef(i) can generally distinguish different fault signals, but different manifold energies corcoef(i) have different abilities. First, corcoef(1) can generally distinguish four rolling-bearing failures. corcoef(2) is also suitable for distinguishing failures, except for the normal and ball fault samples. corcoef(3) failed to distinguish the ball fault and outer-race fault, corcoef(4) failed to distinguish the inner-race fault and ball fault, and corcoef(5) and corcoef(6) failed to distinguish all faults. Thus, corcoef(1) is accepted as the parameter that is best able to distinguish different rolling-bearing failures.

Figure 11 presents the energy correlation coefficient (hereafter denoted as "energy corcoef(i)"), which is calculated using six large energy bands of the HHT time-frequency energy histogram without manifold analysis. As shown in Figure 11, the energy corcoef(i), where $i = 1, \ldots, 6$, cannot provide clear distinctions and thus is not suitable for classifying different rolling-bearing faults.

4.3.3. Manifold Energy Sparsity. The energy distributions of different fault signals are different, as are the energy distributions of different regions in the time-frequency energy histogram. Figure 12(a) presents the manifold energy sparsity of four rolling-bearing signals. As shown in Figure 12(a), the manifold energy sparsity can effectively distinguish different fault samples and can be used to classify different rolling-bearing faults. Figure 12(b) presents the energy sparsity of four types of signal samples. Although the energy sparsity can distinguish different rolling-bearing samples, the energy sparsity within the sample fluctuations, and the difference in energy sparsity of the inter-class sample is not obvious. Therefore, the energy sparsity is not a rolling-bearing fault parameter.

4.3.4. Manifold Energy Mutual Information. We divide the tensor manifold time-frequency energy histogram and HHT time-frequency energy histogram of the four types of signal

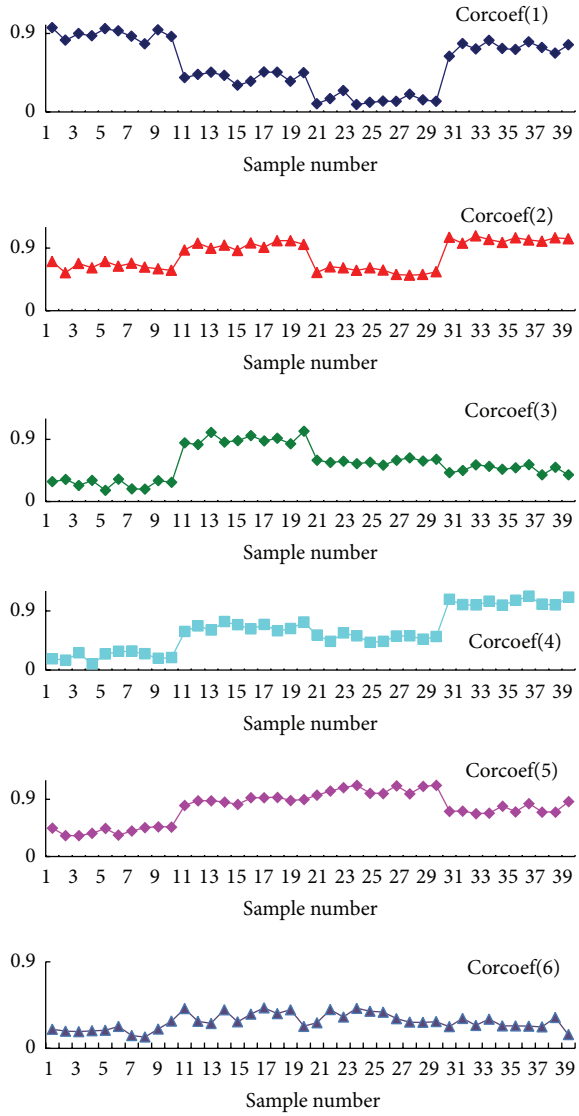


Figure 10: Manifold energy correlation coefficient of the four fault signals.

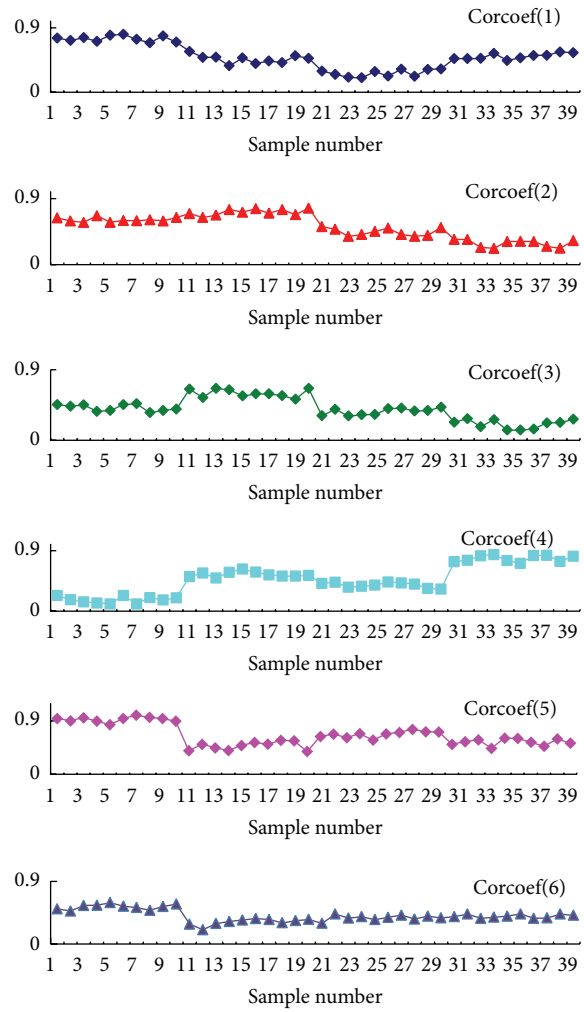


Figure 11: Energy correlation coefficient of the four fault signals.

samples into 6 regions based on the frequency. Then, we calculate the corresponding mutual manifold energy information and mutual energy information.

As described in Figure 13(a), different rolling-bearing faults can be accurately distinguished using the mutual manifold energy information, which is clearly different among the fault samples; thus, the mutual manifold energy information can be used as the rolling-bearing fault characteristic parameter. Figure 13(b) illustrates that the mutual energy information of normal and ball fault samples is similar, and, thus, these two fault types cannot be distinguished. Therefore, the energy mutual information is not suitable for use as the rolling-bearing fault characteristic parameter.

4.3.5. Manifold Energy Kurtosis. We calculate the manifold energy kurtosis and energy kurtosis based on the corresponding energy histograms. Figure 14(a) presents the manifold energy kurtosis of the four different samples; the manifold energy kurtosis varies significantly, and, thus, the mutual energy manifold information can be used to classify different rolling-bearing faults. Figure 14(b) presents the energy kurtosis of the four different samples; as shown, all samples exhibit highly similar energy kurtosis values. In particular, the values of the normal fault, ball fault, and outer-race fault are extremely similar. Therefore, we cannot distinguish different rolling-bearing faults using energy kurtosis.

4.4. Discussion. The merits of the proposed method for extracting the tensor manifold time-frequency characteristic parameters are mainly based on the fact that the tensor manifold time-frequency feature explores the time-varying characteristic of the nonstationary fault signals. The tensor manifold utilizes the HHT time-frequency fault feature of the rolling-bearing fault vibration signals. Thus, the advanced feature is suitable for nonstationary vibration signals. Moreover, there are simple features that are widely used for classification in rolling-bearing fault diagnosis, such as time-domain features (e.g., kurtosis, variance), frequency-domain

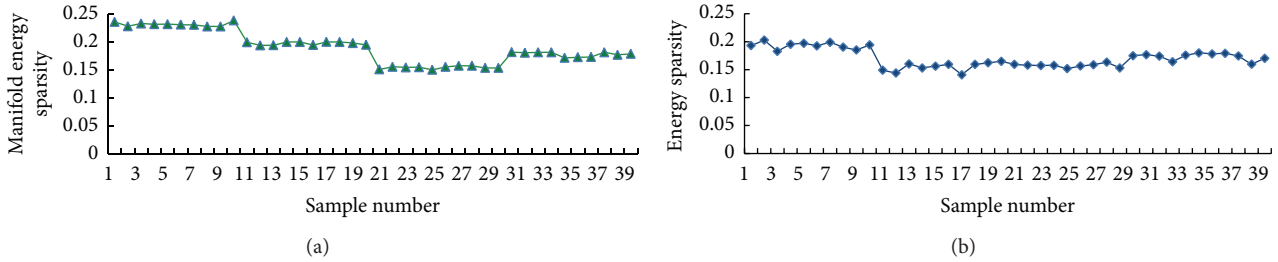


Figure 12: (a) Manifold energy sparsity of the four fault signals and (b) energy sparsity of the four fault signals.

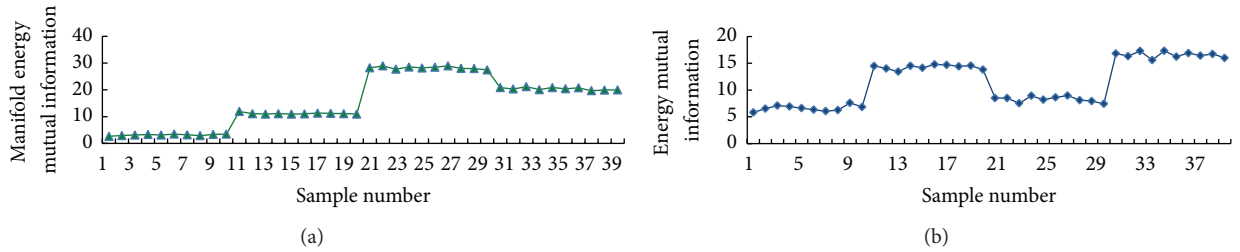


Figure 13: (a) Manifold energy mutual information of the four fault signals and (b) energy mutual information of the four fault signals.

features (e.g., subband energy), and time-frequency domain features (e.g., HHT time-frequency spectrum). These simple features are not as advantageous as the tensor manifold time-frequency features for capturing synthetic signal information. Thus, we use the tensor manifold time-frequency parameters for rolling-bearing fault diagnosis in this paper.

To demonstrate the benefit of the proposed parameters, simple features based on the HHT time-frequency spectrum are also conducted to analyze the four types of rolling-bearing fault signals. The test results are presented in Figures 9–13. These simple features perform worse in classification than do the tensor-manifold-based features. To avoid possible mistakes in pattern identification, it is necessary to improve the classification capability for reliable pattern diagnosis by exploring advanced features, which is the purpose of this paper.

5. Rolling-Bearing Fault Classification

The preceding analysis demonstrates that the five parameters (manifold energy entropy, manifold energy correlation coefficient, manifold energy sparsity, manifold energy mutual information, and manifold energy kurtosis) can efficiently distinguish the rolling-bearing fault states. Thus, they are used as the PNN input parameters for the bearing fault classification.

To verify the effectiveness of the manifold feature for identifying the four bearing faults, 20 samples of four types (normal, inner-race faults, ball faults, and outer-race faults) were used as training samples. The other 20 samples of each type were used for classification purposes. Each sample was extracted for the five aforementioned manifold feature parameters. The characteristic parameters of 80 training samples were used to train the PNN, and the numbers of nodes in the four PNNs were 5, 30, 4, and 4. Finally, the characteristic parameters of the 80 to-be-classified samples were input into the PNN for classification. The PNN classification results of the four bearing faults are shown in Table 1.

Table 1 illustrates that, when the tensor manifold time-frequency characteristic parameters are used as inputs for the PNN, four types of rolling-bearing fault samples can be effectively distinguished and each of the 20 to-be-classified samples for each type of fault can be correctly classified. The normal sample classification exhibits the best results, whereas the minimum components of the category vectors of the inner-race fault samples, ball fault samples, and outer-race fault samples are 0.92, 0.93, and 0.92, respectively. The classification results of the PNN indicate that the rolling-bearing fault condition can be effectively described using the tensor manifold time-frequency characteristic parameters and that the rolling-bearing fault type can be accurately identified with the PNN.

To compare the proposed method with traditional extraction methods, we extract the five defined parameters of the same training and to-be-classified samples using the traditional HHT time-frequency method as the PNN input parameters for the bearing fault classification. The results are shown in Table 2.

Table 2 illustrates that, when the HHT time-frequency characteristic parameters are used as inputs to the PNN, four types of rolling-bearing fault samples can generally be distinguished, but the distinction is not adequate. The normal sample classification exhibits the best results, whereas the minimum components of the category vector of the inner-race fault samples, ball fault samples, and outer-race faults are 0.69, 0.76, and 0.73, respectively.

In Table 1, the minimum components of the category vectors of the four bearing faults are 0.92. In contrast, in Table 2, except the normal-state, the components of the

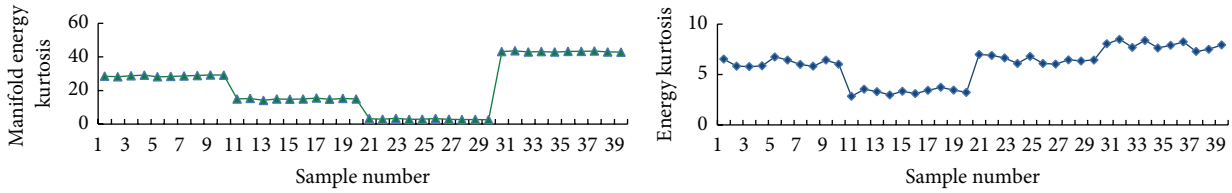


Figure 14: (a) Manifold energy kurtosis of the four fault signals and (b) energy kurtosis of the four fault signals.

Table 1: Classification results of the four bearing faults using the PNN with the proposed tensor manifold features.

Sample	Category vector of the normal-state samples				Category vector of the inner-race fault samples				Category vector of the ball fault samples				Category vector of the outer-race fault samples			
1	1	0	0	0	0	0.97	0.01	0.02	0	0.01	0.99	0	0	0.01	0.02	0.97
2	1	0	0	0	0	0.99	0	0.01	0	0.02	0.97	0.01	0	0.05	0.03	0.92
3	1	0	0	0	0	0.95	0.03	0.02	0	0.02	0.97	0.01	0	0.03	0.02	0.95
4	1	0	0	0	0	0.98	0.01	0.01	0	0.03	0.95	0.02	0	0.02	0.05	0.93
5	1	0	0	0	0	0.93	0.05	0.02	0	0.02	0.98	0	0	0.02	0.02	0.96
6	1	0	0	0	0	0.97	0.02	0.01	0	0.01	0.96	0.03	0	0.05	0.03	0.92
7	1	0	0	0	0	0.95	0.02	0.03	0	0.02	0.97	0.01	0	0.02	0.01	0.97
8	1	0	0	0	0	0.92	0.03	0.05	0	0.03	0.93	0.04	0	0.04	0.03	0.93
9	1	0	0	0	0	0.94	0.02	0.04	0	0.02	0.95	0.03	0	0.03	0.02	0.95
10	1	0	0	0	0	0.96	0.02	0.02	0	0.01	0.97	0.02	0	0.01	0.01	0.98
11	1	0	0	0	0	0.93	0.04	0.03	0	0.01	0.98	0.01	0	0.03	0.01	0.96
12	1	0	0	0	0	0.96	0.03	0.01	0	0.01	0.99	0	0	0.02	0.03	0.95
13	1	0	0	0	0	0.97	0.01	0.02	0	0.02	0.96	0.03	0	0.01	0.02	0.97
14	1	0	0	0	0	0.97	0.01	0.02	0	0.02	0.97	0.01	0	0.01	0.03	0.96
15	1	0	0	0	0	0.95	0.03	0.02	0	0.02	0.95	0.03	0	0.06	0.02	0.92
16	1	0	0	0	0	0.98	0.01	0.01	0	0	0.98	0.02	0	0.03	0.04	0.93
17	1	0	0	0	0	0.96	0.01	0.03	0	0.03	0.96	0.01	0	0.03	0.02	0.95
18	1	0	0	0	0	0.97	0.02	0.01	0	0.01	0.97	0.02	0	0.03	0.01	0.96
19	1	0	0	0	0	0.93	0.03	0.04	0	0.02	0.96	0.02	0	0.03	0.04	0.93
20	1	0	0	0	0	0.96	0.01	0.03	0	0.03	0.93	0.04	0	0.03	0.05	0.92

category vectors of the inner-race fault, the ball fault, and the outer-race fault above 0.85 account for 65%, 80%, and 60%, respectively. Compared with the results of the PNN, which uses tensor manifold time-frequency characteristic parameters as inputs, the classification performance with traditional HHT time-frequency features is relatively poor.

To verify the effectiveness of the manifold feature for identifying different damage degrees in the same fault type of rolling-bearing status, the proposed method was used to classify the inner-race fault samples with different damage degrees. 20 samples of four degrees (normal, mild damage, moderate damage, and severe damage) were used as training samples. The other 10 samples of each degree were used for classification purposes. The PNN classification results of the to-be-classified inner-race fault samples are shown in Table 3.

To compare the proposed method with traditional extraction methods, we extract the five defined parameters of the same training and to-be-classified inner-race fault samples with different damage degrees using the traditional HHT time-frequency method as the PNN input parameters for the bearing fault classification. The results are shown in Table 4.

Table 4 depicts that, when the HHT time-frequency characteristic parameters are used as inputs to the PNN, inner-race fault samples with different damage degrees can generally be distinguished, but the distinction is not adequate. The minimum components of the category vector of the inner-race mild-damage fault samples, moderate-damage fault samples, and severe-damage fault samples are 0.83, 0.86, and 0.78, respectively.

In Table 3, the minimum components of the category vectors of the four bearing faults are 0.93. In contrast, in Table 4, except the normal-state, the components of the category vectors of the inner-race fault, the ball fault, and the outer-race fault above 0.85 account for 90%, 100%, and 60%, respectively. Compared with the results in Table 3, the results of the PNN, which uses the traditional HHT time-frequency characteristic parameters as inputs, indicate a relatively poor classification performance.

Table 2: Classification results of the four bearing faults using the PNN with traditional HHT time-frequency features.

Sample	Category vector of the normal-state samples				Category vector of the inner-race fault samples				Category vector of the ball fault samples				Category vector of the outer-race fault samples			
1	1	0	0	0	0	0.90	0.05	0.05	0	0.10	0.89	0.01	0	0.06	0.07	0.83
2	1	0	0	0	0	0.92	0.03	0.05	0	0.12	0.87	0.01	0	0.05	0.06	0.89
3	1	0	0	0	0	0.89	0.05	0.06	0	0.12	0.79	0.09	0	0.07	0.08	0.85
4	1	0	0	0	0	0.90	0.06	0.04	0	0.13	0.85	0.12	0	0.05	0.04	0.91
5	1	0	0	0	0	0.87	0.06	0.07	0	0.10	0.78	0.12	0	0.06	0.08	0.86
6	1	0	0	0	0	0.92	0.04	0.08	0	0.13	0.76	0.11	0	0.15	0.06	0.79
7	1	0	0	0	0	0.91	0.05	0.04	0	0.07	0.87	0.06	0	0.06	0.07	0.87
8	1	0	0	0	0	0.82	0.08	0.1	0	0.02	0.93	0.05	0	0.15	0.12	0.73
9	1	0	0	0	0	0.87	0.08	0.05	0	0.08	0.85	0.07	0	0.07	0.08	0.75
10	1	0	0	0	0	0.93	0.05	0.02	0	0.07	0.87	0.06	0	0.05	0.07	0.88
11	1	0	0	0	0	0.69	0.14	0.17	0	0.01	0.88	0.11	0	0.06	0.08	0.86
12	1	0	0	0	0	0.74	0.12	0.14	0	0.20	0.79	0.01	0	0.07	0.08	0.85
13	1	0	0	0	0	0.80	0.12	0.08	0	0.08	0.88	0.04	0	0.03	0.06	0.91
14	1	0	0	0	0	0.86	0.09	0.05	0	0.07	0.87	0.06	0	0.07	0.04	0.89
15	1	0	0	0	0	0.75	0.23	0.02	0	0.06	0.89	0.05	0	0.08	0.09	0.83
16	1	0	0	0	0	0.78	0.10	0.12	0	0.03	0.93	0.02	0	0.09	0.10	0.81
17	1	0	0	0	0	0.86	0.12	0.12	0	0.06	0.87	0.07	0	0.07	0.04	0.89
18	1	0	0	0	0	0.87	0.06	0.07	0	0.09	0.89	0.02	0	0.03	0.08	0.89
19	1	0	0	0	0	0.83	0.13	0.14	0	0.02	0.91	0.07	0	0.07	0.01	0.83
20	1	0	0	0	0	0.92	0.05	0.03	0	0.05	0.92	0.03	0	0.07	0.11	0.82

Table 3: Classification results of the inner-race fault using the PNN with the proposed tensor manifold features.

Sample	Category vector of the normal-state samples				Category vector of the inner-race mild-damage fault samples				Category vector of the inner-race moderate-damage fault samples				Category vector of the inner-race severe-damage fault samples			
1	1	0	0	0	0	0.97	0.01	0.02	0	0.01	0.98	0.01	0	0.01	0.02	0.97
2	1	0	0	0	0	0.96	0.02	0.02	0	0.02	0.97	0.01	0	0.02	0.02	0.96
3	1	0	0	0	0	0.98	0.01	0.01	0	0.01	0.97	0.02	0	0.03	0.02	0.95
4	1	0	0	0	0	0.95	0.02	0.03	0	0.03	0.95	0.02	0	0.02	0.02	0.96
5	1	0	0	0	0	0.96	0.02	0.02	0	0.01	0.98	0.01	0	0.01	0.03	0.96
6	1	0	0	0	0	0.97	0.01	0.02	0	0.01	0.96	0.03	0	0.05	0.02	0.93
7	1	0	0	0	0	0.94	0.05	0.02	0	0.02	0.95	0.03	0	0.03	0.02	0.95
8	1	0	0	0	0	0.95	0.03	0.02	0	0.02	0.97	0.01	0	0.01	0.01	0.98
9	1	0	0	0	0	0.94	0.02	0.04	0	0.04	0.94	0.02	0	0.03	0.03	0.94
10	1	0	0	0	0	0.93	0.02	0.05	0	0.04	0.93	0.03	0	0.03	0.01	0.96

6. Conclusion

This paper studies the problem of rolling-bearing fault feature extraction. A time-frequency feature extraction method based on tensor manifolds for rolling bearings was proposed to overcome the deficiencies of the traditional HHT time-frequency feature extraction methods and to remove redundant time-frequency feature information. The HHT time-frequency energy histograms of the rolling-bearing fault signal were used to compose high-dimensional time-frequency fault feature sets. On this basis, the signal time-frequency characteristics were extracted using tensor manifold learning. Five tensor manifold time-frequency characteristic parameters were defined: manifold energy entropy, manifold energy correlation coefficient, manifold energy sparsity, manifold energy mutual information, and manifold energy kurtosis. These characteristic parameters and a PNN were combined to accurately classify rolling-bearing fault samples. The tensor manifold method can realize the nonlinear fusion of the time-frequency information, which can effectively extract the intrinsic nonlinear characteristics of high-dimensional time-frequency combination, and avoid the loss of information

Table 4: Classification results of the inner-race fault using the PNN with traditional HHT time-frequency features.

Sample	Category vector of the normal-state samples					Category vector of the inner-race mild-damage fault samples			Category vector of the inner-race moderate-damage fault samples				Category vector of the inner-race severe-damage fault samples			
1	1	0	0	0	0	0.92	0.05	0.03	0	0.11	0.87	0.02	0	0.12	0.05	0.83
2	1	0	0	0	0	0.87	0.06	0.07	0	0.06	0.89	0.05	0	0.05	0.06	0.89
3	1	0	0	0	0	0.88	0.06	0.06	0	0.07	0.91	0.02	0	0.06	0.06	0.88
4	1	0	0	0	0	0.91	0.07	0.02	0	0.03	0.92	0.05	0	0.06	0.02	0.92
5	1	0	0	0	0	0.83	0.09	0.08	0	0.06	0.90	0.04	0	0.03	0.06	0.91
6	1	0	0	0	0	0.89	0.05	0.06	0	0.09	0.86	0.05	0	0.05	0.06	0.89
7	1	0	0	0	0	0.92	0.04	0.04	0	0.05	0.91	0.04	0	0.06	0.03	0.91
8	1	0	0	0	0	0.90	0.06	0.04	0	0.03	0.92	0.05	0	0.05	0.12	0.83
9	1	0	0	0	0	0.91	0.04	0.05	0	0.09	0.86	0.05	0	0.15	0.07	0.78
10	1	0	0	0	0	0.92	0.03	0.05	0	0.05	0.89	0.06	0	0.11	0.06	0.83

caused by traditional manifold-learning methods. Compared with the HHT time-frequency characteristic parameters, the tensor manifold time-frequency characteristic parameters can more effectively distinguish the four bearing faults, different damage degrees in the same fault type, and different fault types with different damage degrees because of its strong nonlinearity and reduced information redundancy. The effectiveness of the proposed method was verified using real rolling-bearing fault signals. Thus, this paper provides an important method to solve the rolling-bearing feature extraction problems.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (no. 51375067), the Aviation Science Foundation of China (no. 20132163010), and the Fundamental Research Funds for the Central Universities of China (no. DUT13JS08).

References

[1] Q. Liu, F. Chen, Z. Zhou, and Q. Wei, "Fault diagnosis of rolling bearing based on wavelet package transform and ensemble empirical mode decomposition," *Advances in Mechanical Engineering*, vol. 2013, Article ID 792584, 6 pages, 2013.

[2] X. S. Lou and K. A. Loparo, "Bearing fault diagnosis based on wavelet transform and fuzzy inference," *Mechanical Systems and Signal Processing*, vol. 18, no. 5, pp. 1077–1095, 2004.

[3] F. Cong, J. Chen, G. Dong, and M. Pecht, "Vibration model of rolling element bearings in a rotor-bearing system for fault diagnosis," *Journal of Sound and Vibration*, vol. 332, no. 8, pp. 2081–2097, 2013.

[4] L. Jiang, J. Xuan, and T. Shi, "Feature extraction based on semi-supervised kernel Marginal Fisher analysis and its application in bearing fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 41, no. 1-2, pp. 113–126, 2013.

[5] Y. Lei, Z. He, and Y. Zi, "A new approach to intelligent fault diagnosis of rotating machinery," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1593–1600, 2008.

[6] Z. K. Peng, P. W. Tse, and F. L. Chu, "A comparison study of improved Hilbert-Huang transform and wavelet transform: application to fault diagnosis for rolling bearing," *Mechanical Systems and Signal Processing*, vol. 19, no. 5, pp. 974–988, 2005.

[7] A. Papandreou-Suppappola, *Applications in TimeFrequency Signal Processing*, vol. 10, CRC Press, Boca Raton, Fla, USA, 2003.

[8] B. Boashash, *Time Frequency Signal Analysis and Processing*, Prentice Hall, New York, NY, USA, 2002.

[9] K. Gröchenig, *Foundations of Time–Frequency Analysis*, Birkhauser, Boston, Mass, USA, 2000.

[10] S. Stanković, "Time-frequency analysis and its application in digital watermarking," *Eurasip Journal on Advances in Signal Processing*, vol. 2010, Article ID 579295, 2010.

[11] Z. W. Wang and C. Y. Hu, "Shock spectra and damage boundary curves for nonlinear package cushioning system," *Packaging Technology and Science*, vol. 12, no. 5, pp. 207–217, 1999.

[12] J. Y. Zhang, Y. Y. Zhang, and Y. B. Xie, "Applications of time-frequency analysis based on wavelet packet to incipient impulse faults diagnosis," *Journal of Vibration Engineering*, vol. 13, no. 2, pp. 66–72, 2000.

[13] L. M. Zhu, X. W. Niu, B. L. Zhong, and H. Ding, "Approach for extracting the time-frequency feature of a signal with application to machine condition monitoring," *Journal of Vibration Engineering*, vol. 17, no. 4, pp. 71–76, 2004.

[14] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London A*, vol. 454, no. 1971, pp. 903–995, 1998.

[15] M. Gandetto, M. Guainazzo, and C. S. Regazzoni, "Use of time-frequency analysis and neural networks for mode identification in a wireless software-defined radio approach," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 12, pp. 1778–1790, 2004.

[16] Y. G. Lei, Z. J. He, and Y. Y. Zi, "EEMD method and WNN for fault diagnosis of locomotive roller bearings," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7334–7341, 2011.

[17] P. Shan and M. Li, "Nonlinear time-varying spectral analysis: HHT and MODWPT," *Mathematical Problems in Engineering*, vol. 2010, Article ID 618231, 14 pages, 2010.

[18] H. Li, P. Zhou, and Z. Zhang, "An investigation into machine pattern recognition based on time-frequency image feature extraction using a support vector machine," *Journal of Mechanical Engineering Science*, vol. 224, no. 4, pp. 981–994, 2010.

[19] H. K. Li, S. Zhou, and W. Z. Huang, "Time-frequency image feature extraction for machine condition classification and its application," *Journal of Vibration and Shock*, vol. 29, no. 7, pp. 184–188, 2010.

[20] Q. He, "Time-frequency manifold for nonlinear feature extraction in machinery fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 35, no. 1-2, pp. 200–218, 2013.

[21] F. Dornaika and A. Assoum, "Enhanced and parameterless locality preserving projections for face recognition," *Neurocomputing*, vol. 99, pp. 448–457, 2013.

[22] M. Ravishankar and D. R. Rameshbabu, "Ten-LoPP: tensor locality preserving projections approach for moving object detection and tracking," in *Proceedings of the 9th International Conference on Computing and InformationTechnology*, pp. 291–300, Springer, Berlin, Germany, 2013.

[23] "Bearings Vibration Data Set, Case Western Reserve University," http://csegroups.case.edu/bearingdatacenter/home.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 713678, 9 pages
http://dx.doi.org/10.1155/2014/713678

Research Article

An Analytical Model for Fatigue Crack Propagation Prediction with Overload Effect

Shan Jiang, Wei Zhang, Xiaoyang Li, and Fuqiang Sun

Science and Technology on Reliability and Environmental Engineering Laboratory, School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China

Correspondence should be addressed to Wei Zhang; zhangwei.dse@buaa.edu.cn

Received 2 May 2014; Revised 22 June 2014; Accepted 23 June 2014; Published 23 July 2014

Academic Editor: Xuefeng Chen

In this paper a theoretical model was developed to predict the fatigue crack growth behavior under the constant amplitude loading with single overload. In the proposed model, crack growth retardation was accounted for by using crack closure and plastic zone. The virtual crack annealing model modified by Bauschinger effect was used to calculate the crack closure level in the outside of retardation effect region. And the Dugdale plastic zone model was employed to estimate the size of retardation effect region. A sophisticated equation was developed to calculate the crack closure variation during the retardation area. Model validation was performed in D16 aluminum alloy and 350WT steel specimens subjected to constant amplitude load with single or multiple overloads. The predictions of the proposed model were contrasted with experimental data, and fairly good agreements were observed.

1. Introduction

The damage tolerance concept is widely used in modern aircraft design to ensure flight safety, which has made the prediction of fatigue crack propagation lives of aircraft components under service loading necessary [1, 2]. In 1860s, Paris and Erdogan [3] proposed a fracture mechanics based method for fatigue life prediction, which correlated the fatigue crack growth rate to the applied stress intensity factor range. One issue of Paris' model is that the stress ratio effect is not considered. Many modifications of Paris' law have been proposed in the literatures [4–6], one of the most important modifications being the inclusion of the crack closure concept. Crack closure was first introduced into the fatigue crack growth analysis by Wolf [4]. And then, plenty of research has been done concerning the crack closure using experimental investigation, numerical analysis, and theoretical studies [7–16]. Zhang and Liu [7, 8] performed a state-of-the-art in situ SEM testing to investigate fatigue crack growth behavior continuously within a load cycle. In their study, crack closure's existence and its influence on crack propagation was directly observed. Newman [9, 10] used a strip yield model to analyze the crack closure problems. Budiansky and Hutchinson [13] proposed a method to estimate the crack closure caused by plasticity in plane stress cases under constant cyclic loading. The crack closure models reviewed above can eliminate stress ratio R effect and describe fatigue crack behavior under constant amplitude loading well. However, since crack closure calculation involves the highly nonlinear analysis of cyclic plasticity and contact analysis, direct tracking of crack closure under a variable amplitude loading is extremely difficult and of high computational cost. So, the plastic zone concept is also employed to describe the crack growth behavior under general loading conditions. Willenborg et al. and Wheeler [17, 18] proposed a series of models to analyze the fatigue crack growth under variable loading cases. In their papers, the plastic zone concept was used to explain the overload retardation effect. Plentiful literatures were presented to investigate fatigue crack growth behavior with the single overload [19–22], and many modifications have been proposed to correlate experimental observations. However, in most of these models, empirical coefficients are added to match the testing data under the influence of retardation, instead of modeling directly based on the mechanisms, such as the crack closure. Moreover, the recent in situ SEM experimental observations have shown that

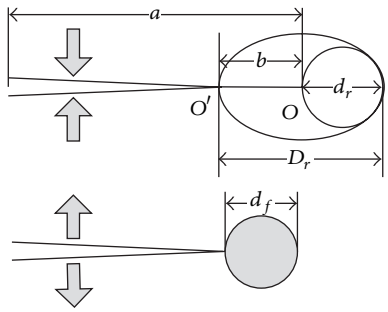


Figure 1: Schematic illustration of real crack and virtual crack model.

single overload has a significant impact on the crack closure within the large plastic zone [23]. Therefore, crack closure variation is probably a good explanation of the retardation phenomenon. In this paper, a mechanism-based method will be developed to predict fatigue crack growth behavior under the single overload, in which crack closure and plastic zone concept are considered.

The paper is organized as follows. First, the fatigue crack growth model and the modified virtual crack annealing crack closure model are briefly reviewed. Second, our proposed model for crack growth retardation behavior estimation is discussed in detail. The model is derived based on plastic zone and crack closure variation. And then model validation is performed using the experimental data in D16 aluminum alloy and 350WT steel from the literature. Finally, some conclusions and future work are given based on the current study.

2. Methodology Development

A general fatigue crack growth model can be expressed as (1) by Wolf [4]. The fatigue crack growth rate is determined by the effective stress intensity factor range

$$\frac{da}{dN} = C\left(\Delta K_{\text{eff}}\right)^m, \qquad \Delta K_{\text{eff}} = K_{\max} - K_{\text{op}}, \tag{1}$$

where da/dN is the crack growth rate, ΔK_{eff} is the effective stress intensity factor, $K_{\max}$ is the stress intensity factor of the peak load, and K_{op} is the crack closure level; C, m are calibration parameters.

In this paper the virtual crack annealing model is employed to avoid considering the complex contact of crack closure [7]. This analytical closure model is based on plasticity ahead of crack tip, but it does not consider the Bauschinger effect. So a brief derivation of the modified closure model is discussed in the following part.

As shown in Figure 1, in the process of unloading, when the stress level decreases from $\sigma_{\max}$ to σ_{op}, a reversed plastic zone with diameter d_r appears ahead of the real crack tip "O". And as the stress level continues reducing from σ_{op} to $\sigma_{\min}$, the crack tip will fully close and the reversed plastic zone remains constant due to the crack closure. According to the assumption of crack annealing, the closed part that has the length of "b" can be considered as nonfractured material. So, there is an equivalent reversed plastic zone with diameter D_r, and the virtual crack tip position is O'. In the following loading process, when the stress level increases from $\sigma_{\min}$ to σ_{op}, the superposition of the stress within the forward plastic zone with diameter d_f and the residual stress within the distance "b" ahead of the virtual crack is zero and the crack tip opens. So, the equation $d_f = D_r - d_r$ can be established. Additionally, for anisotropic materials, the Bauschinger effect has a significant impact on the cyclic plastic deformation ahead of crack tip. Thus, equation can be expressed as

$$\frac{\pi}{8}\left(\frac{\sigma_{\max}-\sigma_{\min}}{2\beta\sigma_y}\right)^2\pi(a-b) - \frac{\pi}{8}\left(\frac{\sigma_{\max}-\sigma_{\text{op}}}{2\beta\sigma_y}\right)^2\pi(a) = \frac{\pi}{8}\left(\frac{\sigma_{op}-\sigma_{\min}}{\alpha\sigma_y}\right)^2\pi(a-b), \tag{2}$$

where σ_y is tensile yield strength, $\sigma_{\max}$ is maximum stress level, $\sigma_{\min}$ is minimum stress level, σ_{op} is stress level of crack opening, α is the Bauschinger effect factor in loading process, and β is the Bauschinger effect in unloading process.

Since the crack overlapping length is very small compared with the true crack length (i.e., $b \ll a$), its effect on the stress intensity factor calculation can be ignored. Equation (2) can be rewritten as

$$\begin{aligned}&\left(\alpha+4\beta^2\right)\sigma_{\text{op}}^2 - \left(8\beta^2\sigma_{\min}+2\alpha^2\sigma_{\max}\right)\sigma_{\text{op}} + \left(4\beta^2\sigma_{\min}^2+2\alpha^2\sigma_{\max}\sigma_{\min}-\alpha^2\sigma_{\min}^2\right)=0,\\ &\left(1+4\gamma^2\right)\sigma_{\text{op}}^2 - \left(8\gamma^2\sigma_{\min}+2\sigma_{\max}\right)\sigma_{\text{op}} + \left(4\gamma^2\sigma_{\min}^2+2\sigma_{\max}\sigma_{\min}-\sigma_{\min}^2\right)=0,\end{aligned} \tag{3}$$

where γ is equal to β/α. Equation (3) is the general solution using the virtual crack annealing model for the crack opening stress calculation under constant amplitude loadings. Equation (3) can be further simplified as

$$\left(\sigma_{\text{op}}-\sigma_{\min}\right)\left[\left(1+4\gamma^2\right)\sigma_{\text{op}} - 2\sigma_{\max} - \left(4\gamma^2-1\right)\sigma_{\min}\right] = 0. \tag{4}$$

There are two possible solutions for the opening stress level under the proposed virtual crack annealing model. One

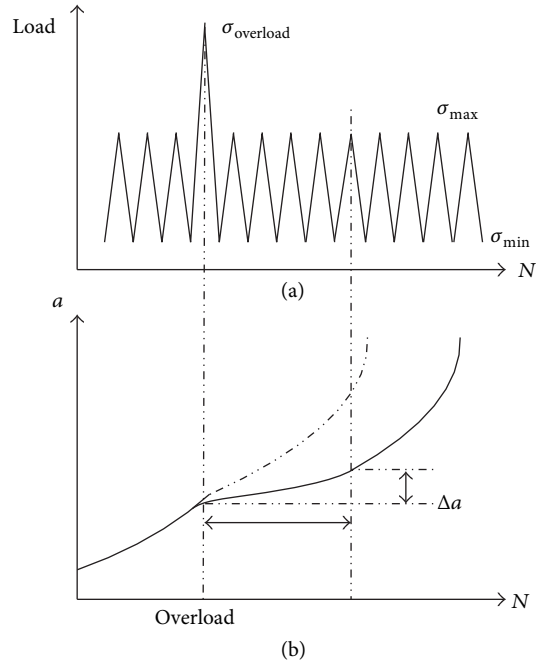


Figure 2: Schematic illustration of CA load interspersed with single-cycle tensile overload.

solution is $\sigma_{op} = \sigma_{min}$, which indicates that there is no crack closure and the overlapped length is zero. The other solution is $\sigma_{op}/\sigma_{max} = (1/(1 + 4\gamma^2))[2 + (4\gamma^2 - 1)R]$. This observation indicates that either there is no crack closure or there is a unique crack closure level under constant amplitude loadings. In this paper, the modified virtual crack annealing model is used to make predictions of the fatigue crack growth behavior in the outside of retardation effect region.

Retardation caused by overloads is a typical phenomenon of loading interaction effect. A great number of papers were presented to investigate the single overload [24–28]. Generally, overloads will induce large plastic deformation ahead of the crack tip and retard the crack growth rate in the subsequence loading of a certain range. As the developing of the crack, the retardation effect gradually recedes. When the crack grows beyond the large plastic zone, the retardation effect vanishes completely. Before the crack growth rate decreases, there is an accelerated crack growing stage, which has been termed delayed retardation. That transient acceleration right after overload is small enough to be neglected in this investigation. As it is shown in Figure 2, the crack growth length is "Δa" under the influence of a single overload.

The crack growth behavior is shown in Figure 3. After the overload, the crack growth rate increases transitorily and then decreases sharply. Once the da/dN value reaches to minimum, crack growth rate increases gradually and finally recovers to the equilibrium constant amplitude loading growth rate.

The crack tip plastic zone associated with the application of a single overload is shown in Figure 4(a). A new approach to estimation of crack opening stress under the influence of single overload is proposed, as shown in Figure 4(b). When

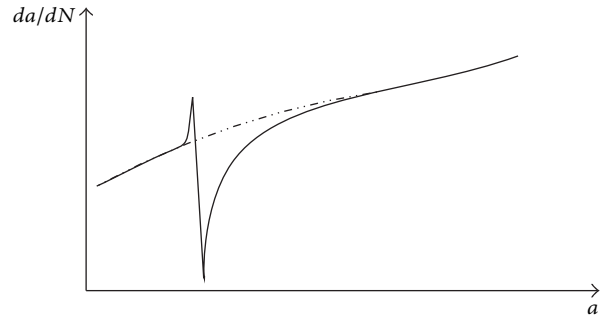


Figure 3: Schematic of delayed retardation due to single tensile overload.

the crack grows to point O_1, a single overload stress is applied and a large plastic zone appears ahead of the crack tip. As the crack penetrates into this plastic zone and grows to point O_3, the retardation effect gradually recedes until it vanishes.

A function of σ_{op}/σ_{max} and the crack length is established to calculate the crack closure variation during the retardation region, which is shown in Figure 4(b). The points C and D in this figure, respectively, represent the situation of crack growing to points O_1 and O_3 in Figure 4(a). Equation (5) can be obtained based on the modified virtual crack annealing model above:

$$Y_C = \left(\frac{\sigma_{op}}{\sigma_{max}}\right)_C = \frac{2}{1+4\gamma^2} + \frac{4\gamma^2 - 1}{1+4\gamma^2} \times \frac{R}{R_{ol}}, \quad (5)$$

where γ is Bauschinger effect factor, R is the load ratio, and R_{ol} is the overload ratio. The left side of (5) is the ordinate of point C. Similarly, the ordinate of point D can be illustrated by the following:

$$Y_D = \frac{2}{1+4\gamma^2} + \frac{4\gamma^2 - 1}{1+4\gamma^2} \times R. \quad (6)$$

The horizontal ordinate of point D is

$$X_D = a_0 + D_m, \quad (7)$$

where a_0 is the horizontal ordinate of point C, which is the crack length when single overload is applied. D_m is the diameter of plastic zone created by that single overload. If point D is considered to be the ordinate origin, the equation of the curve between points C and D is

$$f(x) = \frac{Y_C - Y_D}{(X_C)^n} \cdot (-x)^n, \quad (8)$$

where Y_C and Y_D are, respectively, the ordinate of points C and D and X_C is the horizontal ordinate of point C. When the value of n can be 2, 1, and 0.5, the curve is, respectively, shown as (A), (B), and (C) in Figure 4. Based on the current study, $n = 1$ is a good approximation for aluminum alloy. If the O is considered to be the ordinate origin, (7) can be rewritten as

$$f(x) = Y_D + \frac{Y_C - Y_D}{(X_C - X_D)^n}\left((X_C + D_m) - x\right)^n. \quad (9)$$

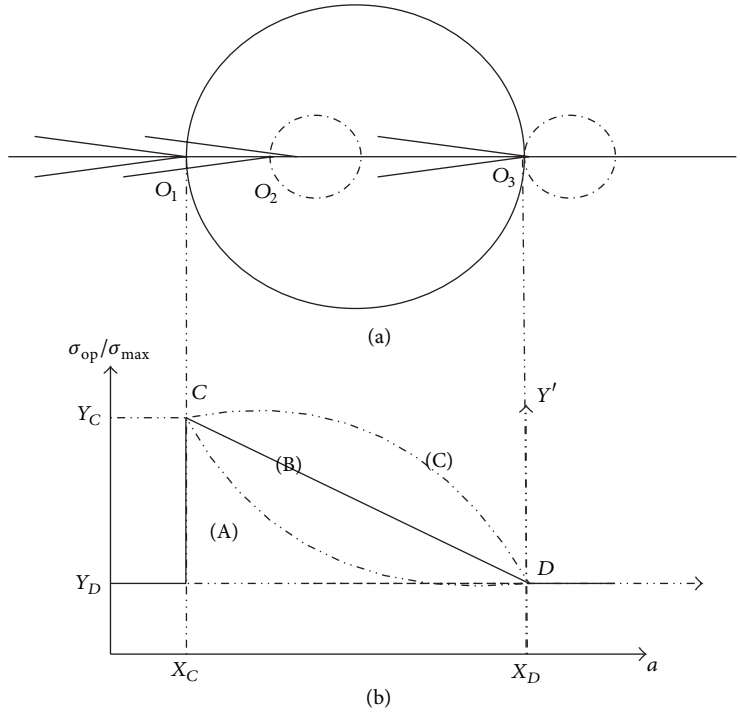


Figure 4: (a) Crack-tip plastic zone due to single tensile overload; (b) the mathematical derivation of σ_{op} based on crack closure model.

Based on the discussion above, the new crack closure model is established as (10), which is used to describe fatigue crack propagation behavior under single overload interspersed in a constant amplitude loading spectrum. Consider

$$\sigma_{op} = \begin{cases} \dfrac{1}{4\gamma^2+1}\left(\left(4\gamma^2-1\right)\sigma_{min}+2\sigma_{max}\right) \\ \qquad \text{before single overload} \\ \dfrac{1}{4\gamma^2+1}\left(\left(4\gamma^2-1\right)\sigma_{min}+2\sigma_{ol}\right) \\ \qquad \text{single overload} \\ \left(\dfrac{2}{1+4\gamma^2}+\dfrac{4\gamma^2-1}{1+4\gamma^2}\times R\right) \\ \quad +\dfrac{\left(\left(4\gamma^2-1\right)/\left(1+4\gamma^2\right)\right)\left(\left(R/R_{ol}\right)-R\right)}{\left(-D_m\right)^n} \\ \quad \times\left(a_0+D_m-a\right)^n \\ \qquad \text{under influence of single overload} \\ \dfrac{1}{4\gamma^2+1}\left(\left(4\gamma^2-1\right)\sigma_{min}+2\sigma_{max}\right) \\ \qquad \text{after influence of single overload.} \end{cases} \tag{10}$$

The cycle by cycle algorithm is used to implement the above model. The flow chart for fatigue crack growth prediction based on the modified crack closure model is shown in Figure 5.

Table 1: Standard chemical composition [28].

Chemical composition of D16 aluminum alloy						
Element	Cu	Mg	Mn	Si	Fe	Zn
Weight %	3.8–4.9	1.2–1.8	0.3–0.9	0.5	0.5	0.3

3. Experimental Validation

3.1. Crack Growth Predictions in D16 under Constant Loading and Single Overload. The experimental data for model validation are from the literature [27, 28]. The material used in both papers was D16 aluminum alloy. The standard chemical composition is shown in Table 1. And the mechanical properties of this material in longitudinal (LT) orientation were as follows: $\sigma_y = 347$ MPa, $K_c = 40 \sim 45$ MPa$*$m$^{0.5}$, and percentage elongation = 12%.

Before any predictions can be made, there are several unknown parameters (see (1) and (10)) that need to be calibrated. Three sets of $da/dN \sim dK$ testing data are used to evaluate these parameters (Figure 6), which are assumed to depend on the material only [27]. The calibrated parameters are $m = 2.4534$ and $C = 1.2331$E-9, and Bauschinger effect factor γ is 0.97.

Then additional two sets of a-N data are used to validate the proposed model. The specimen configuration for these tests is [27] width = 100 mm, length = 500 mm, and thickness = 4 mm. Initial half crack length is $a = 5.0$ mm. The specimen was subjected to a constant amplitude load sequence with $\sigma_{max} = 64$ MPa, $R = 0$. And the overload ratio is 2.

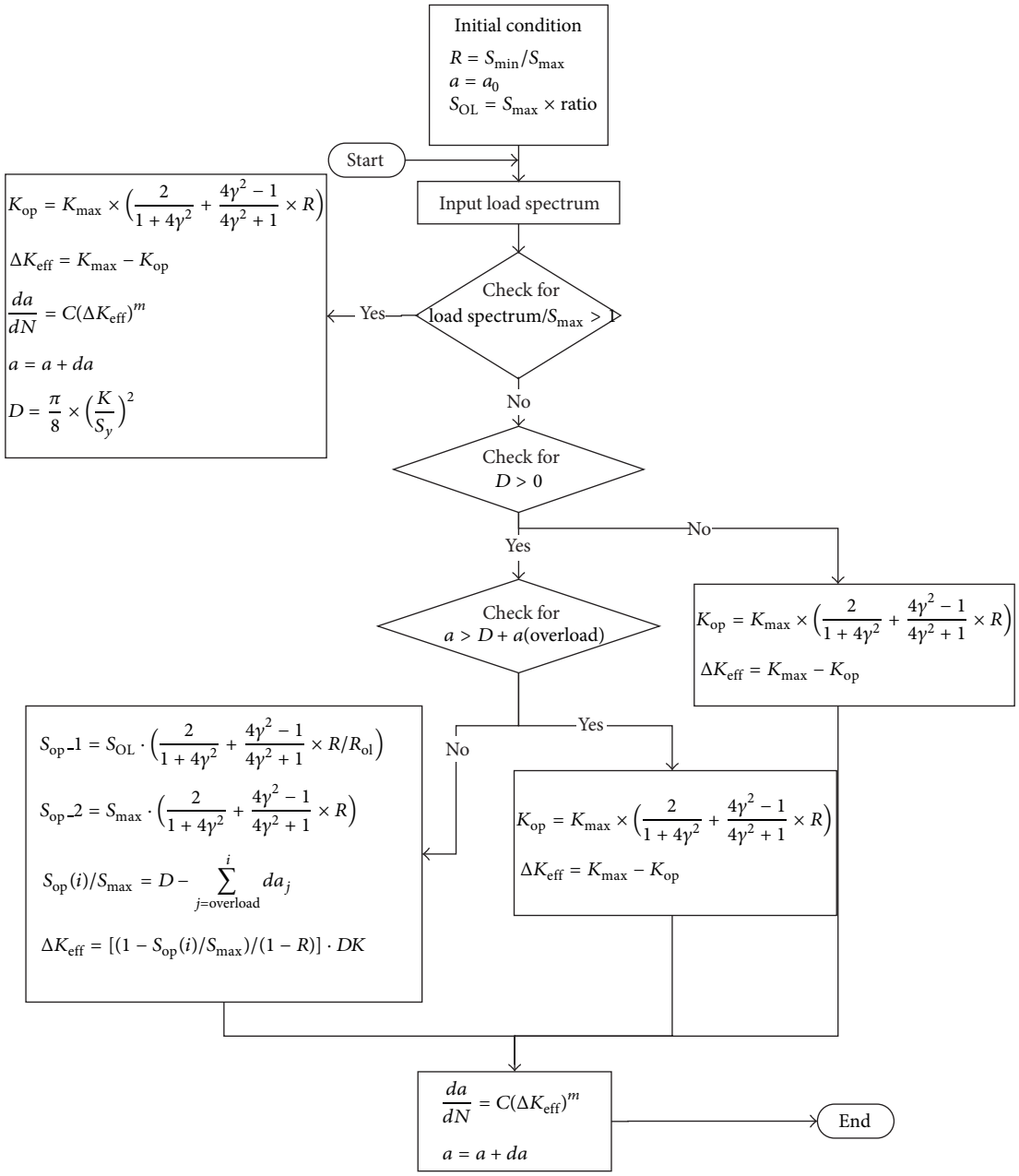


Figure 5: Flow chart for fatigue crack growth calculation.

The model predictions are compared to the experimental data in Figure 7. The x-axis is the number of cycles, and the y-axis is crack length. The red circlets represent the fatigue crack growth under the constant amplitude loading; and the small blue squares are the counterpart under the constant amplitude load with single overload. It is observed that the predictions by proposed model match the testing data well.

More detailed information can be obtained in the da/dN-a curve in Figure 8. Right after the single overload, the fatigue crack growth rate increases in a short time and then decreases sharply. After a certain crack length, the retardation phenomenon vanishes and the crack growth rate turns back to the original trend. Note that very good agreement is observed between the model prediction and experimental data.

Additional set of testing data is used for the model validation in the same material and similar specimen configuration [28]. Specimen dimensions are as follows: length = 180 mm, width = 45 mm, and thickness = 1.5 mm. Initial crack length is $a = 4.0$ mm. The specimens are subjected to two different loading spectra: (1) constant amplitude loading with no overloads ($\sigma_{max} = 60$ MPa, $R = 0.1$) and (2) constant amplitude load with single overload applied at the certain crack length (overload ratio is 2) [28]. Following the same procedure, the comparisons between predictions and experimental data are shown in Figure 9. Under the

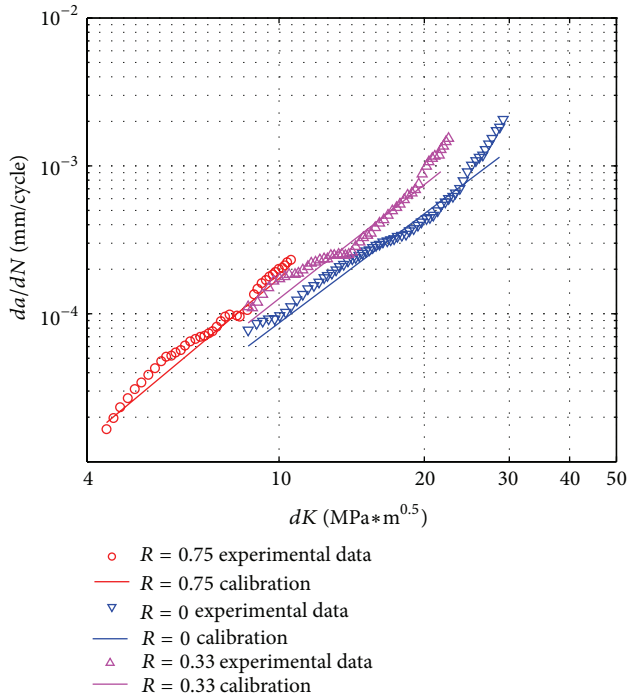


Figure 6: da/dN-dK calibration.

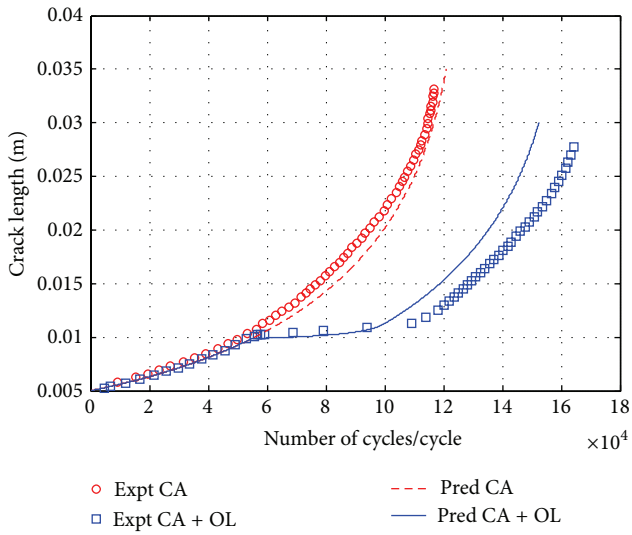


Figure 7: a-N curve for constant amplitude and single overload cases.

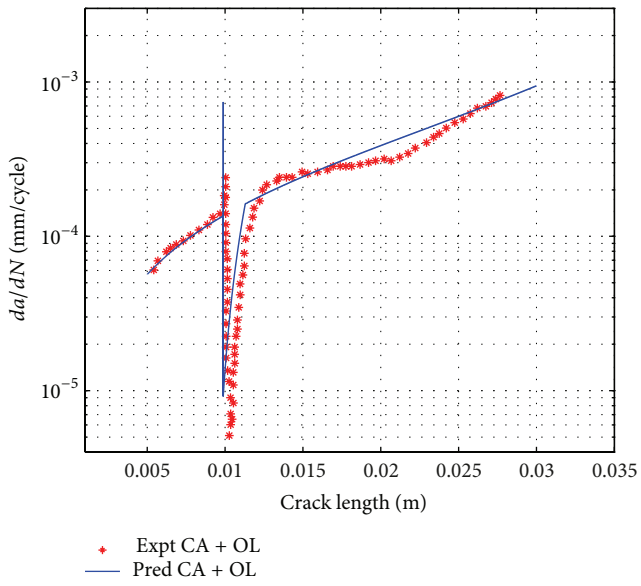


Figure 8: da/dN-a curve for constant amplitude and single overload cases.

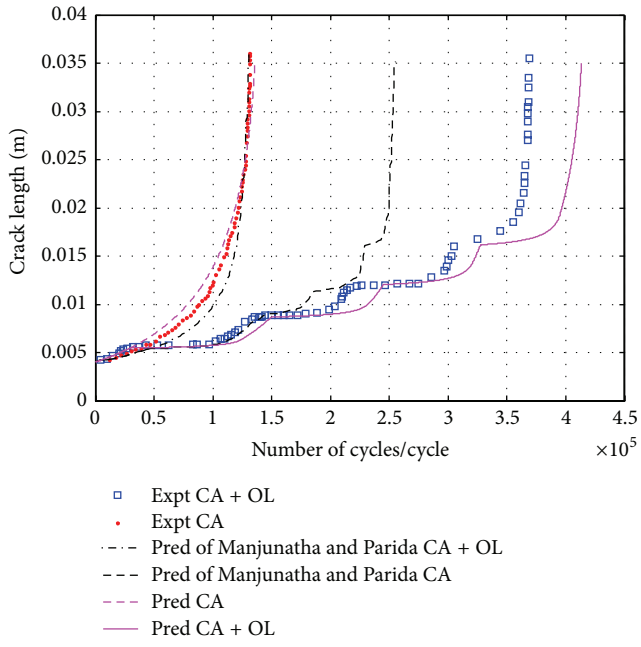


Figure 9: a-N periodic spike loading.

former loading spectrum, both of our proposed model and Manjunatha's model can give very good predictions [28]. In the latter loading case, the single overloads are applied at the crack length of 5.5, 9, 12, and 16 mm, respectively. The prediction of our proposed model is just slightly slower than the experimental data, while Manjunatha's prediction is much faster than that of the experimental data. The fatigue life ratios expressed as $N_{\text{expt}}/N_{\text{pred}}$ are, respectively, calculated to be 0.89 by using the proposed model and 1.44 by using Manjunatha's model. Obviously, the proposed closure based model can give the better predictions.

3.2. Crack Growth Prediction in 350WT under Constant Loading and Single Overload. In this section, the model validation will extend to 350WT steel. The experimental data are given by Taheri et al. [29]. The specimen dimensions are as follows: length = 300 mm, width = 100 mm, and thickness = 5 mm. Initial center crack length is $2a$ = 20.0 mm. Recording test data was started at crack length of 11.17 mm. In addition,

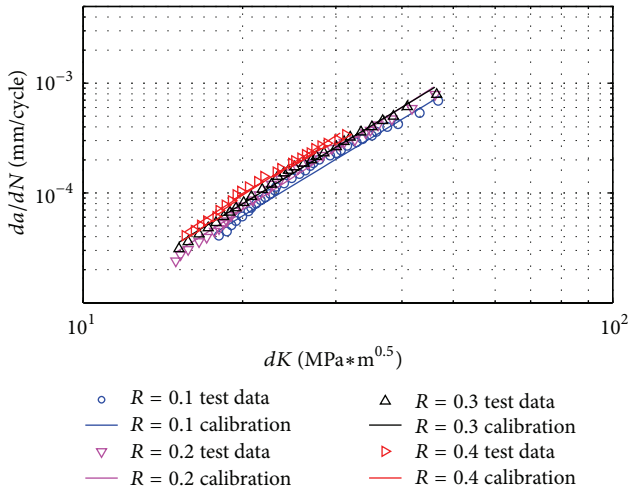


FIGURE 10: da/dN-dK calibration figure.

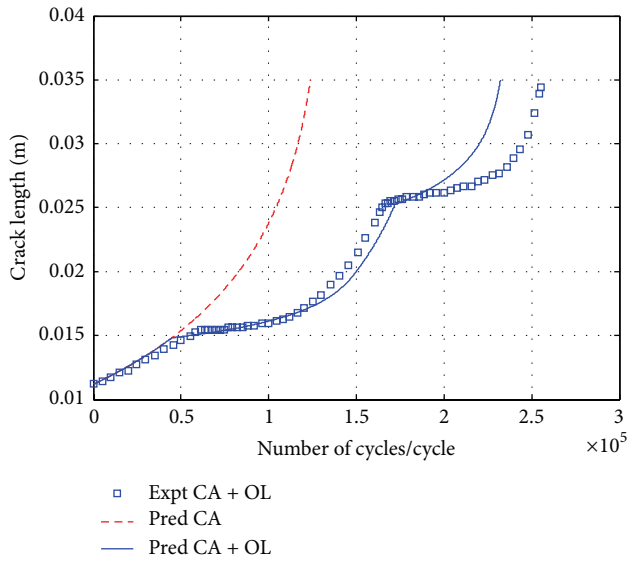


FIGURE 11: a-N curve for constant amplitude and single overload cases.

the mechanical properties of 350WT in longitudinal (LT) orientation are as follows: σ_y = 350 MPa, σ_{UTS} = 524 MPa, Modulus of elasticity = 205 GPa, and Poisson's ration = 0.30 [30].

Similarly, the parameters are calibrated by the $da/dN \sim \Delta K$ data in the literature [19], as shown in Figure 10. The fitting coefficients are $C = 4.3136E - 11$ and $m = 2.9245$, and the Bauschinger effect factor γ is 1.05.

Figure 11 shows the comparison between the predicted a-N curve and the experimental data in 350WT steel specimens. The single overloads are applied at the crack length of 15 and 25 mm. It can be seen that there are two obvious retarded regions right after the overload in the experimental data. The prediction follows almost the same trend as the experimental observation. Predicted total fatigue lives are slightly lower than the experimental values.

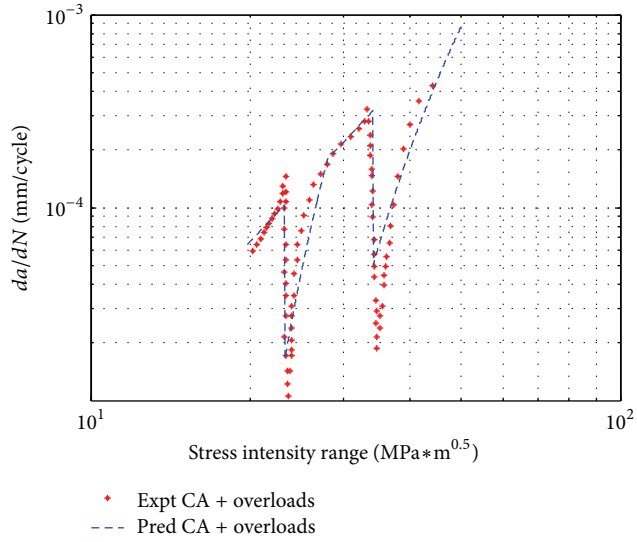


FIGURE 12: ΔK-da/dN cruve for constant amplitude and single overload cases.

In Figure 12, crack growth rate variation caused by overloads is also compared with our model prediction. It is clear that the prediction matches the experimental data very well, especially in the retardation region. Hence, it is proven that the method proposed in this paper is applicable to predict the fatigue crack growth behavior under the constant loading with single overloads in 350WT steel.

4. Conclusion

In this investigation, a theoretical model is developed to predict the fatigue crack growth behavior under the constant amplitude loading with single overload. The crack growth retardation was accounted for by using crack closure concept and plastic zone. Model was validated in D16 aluminum alloy and 350WT steel subjected to several different loading spectra, and the predictions matched experimental data well. The following conclusions can be drawn based on the current investigation.

(1) Fatigue crack growth is slowed down by application of single overload cycle. A convincing reason for this retardation phenomenon is that after the overload a large plastic zone will form ahead of the crack tip, which can increase the crack closure level within this region. And as the crack grows through the large plastic zone, the crack closure level will gradually decrease which can be described as a linear function. The retardation effects disappear after a certain characteristic crack length extension from the overload position. This extension is approximately equal to monotonic plastic zone size caused by the overload.

(2) The proposed model is derived from fatigue crack growth mechanisms (such as crack closure, plastic zone, and Bauschinger effect), and it does not require

any additional parameters which has no physical meaning.

(3) The above statement is only valid under the current investigated loading spectrums and materials. In the future, the whole frame work should be extended to other materials. Additionally, branching and bifurcation caused by overload can also retard the crack growth rate, which should be investigated in the future.

Nomenclature

a:	Crack length
Δa:	Crack growth in one cycle
da:	Infinitesimal crack increment
da/dN:	Fatigue crack growth rate per cycle
$\sigma_{\min}$, $\sigma_{\max}$:	Minimum and maximum stress in one loading cycle
σ_{op}:	Stress level at which the crack begins to grow
σ_{ol}:	Stress level of single overload
R:	Stress ratio
R_{ol}:	Overload ratio
$K_{\max}$, $K_{\min}$:	Maximum/minimum stress intensity factor
ΔK:	Stress intensity factor range
K_{op}:	Stress intensity factor at which the crack begins to grow
ΔK_{eff}:	Effective stress intensity factor range
D_m:	Monotonic plastic zone size
D_f:	Forward plastic zone size
d_r, D_r:	Reverse plastic zone size
σ_y:	Material yield strength
α:	Bauschinger effect factor in loading process
β:	Bauschinger effect factor in unloading process
γ:	Bauschinger effect factor.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The research is financially supported by the specialized research fund for the doctoral program of higher education funding under the contract no. 20131102120047.

References

[1] R. C. Alderliesten and J. J. Homan, "Fatigue and damage tolerance issues of Glare in aircraft structures," *International Journal of Fatigue*, vol. 28, no. 10, pp. 1116–1123, 2006.

[2] A. T. Kermanidis, P. V. Petroyiannis, and S. G. Pantelakis, "Fatigue and damage tolerance behaviour of corroded 2024 T351 aircraft aluminum alloy," *Theoretical and Applied Fracture Mechanics*, vol. 43, no. 1, pp. 121–132, 2005.

[3] P. Paris and F. Erdogan, "A critical analysis of crack propagation laws," *Journal of Fluids Engineering*, vol. 85, no. 4, pp. 528–533, 1963.

[4] E. Wolf, "Fatigue crack closure under cyclic tension," *Engineering Fracture Mechanics*, vol. 2, no. 1, pp. 37–45, 1970.

[5] J. C. Newman Jr., *Prediction of Crack Growth under Variable-Amplitude Loading in Thin-Sheet 2024-T3 Aluminum Alloys*, Engineering against Fatigue, University of Sheffield, 1997.

[6] W. Elbert, "The significance of fatigue crack closure," in *Damage Tolerance in Aircraft Structures: A Symposium Presented at the Seventy-Third Annual Meeting American Society for Testing and Materials*, pp. 486–230, ASTM International, Toronto, Canada, June 1971.

[7] W. Zhang and Y. Liu, "In situ SEM testing for crack closure investigation and virtual crack annealing model development," *International Journal of Fatigue*, vol. 43, pp. 188–196, 2012.

[8] W. Zhang and Y. Liu, "Investigation of incremental fatigue crack growth mechanisms using in situ SEM testing," *International Journal of Fatigue*, vol. 42, pp. 14–23, 2012.

[9] J. C. Newman Jr., "A crack-closure model for predicting fatigue crack growth under aircraft spectrum loading," *ASTM International—STP Series*, vol. 748, pp. 53–84, 1981.

[10] J. C. Newman Jr., "A crack opening stress equation for fatigue crack growth," *International Journal of Fracture*, vol. 24, no. 4, pp. R131–R135, 1984.

[11] J. Z. Zhang and P. Bowen, "On the finite element simulation of three-dimensional semi-circular fatigue crack growth and closure," *Engineering Fracture Mechanics*, vol. 60, no. 3, pp. 341–360, 1998.

[12] F. V. Antunes, A. G. Chegini, R. Branco, and D. Camas, "A numerical study of plasticity induced crack closure under plane strain conditions," *International Journal of Fatigue*, 2014.

[13] B. Budiansky and J. W. Hutchinson, "Analysis of closure in fatigue crack growth," *Journal of Applied Mechanics*, vol. 45, no. 2, pp. 267–276, 1978.

[14] J. Llorca and V. S. Gálvez, "Modelling plasticity-induced fatigue crack closure," *Engineering Fracture Mechanics*, vol. 37, no. 1, pp. 185–196, 1990.

[15] P. F. P. de Matos and D. Nowell, "Modeling fatigue crack closure using dislocation dipoles," 2006.

[16] J. H. Kim and S. B. Lee, "Behavior of plasticity-induced crack closure and roughness-induced crack closure in aluminum alloy," *International Journal of Fatigue*, vol. 23, supplement 1, pp. S247–S251, 2001.

[17] J. Willenborg, R. M. Engle, and H. A. Wood, "A crack growth retardation model using an effective stress concept," Tech. Rep., 1971.

[18] O. E. Wheeler, "Spectrum loading and crack growth," *Journal of Basic Engineering Transactions of the ASME*, vol. 94, no. 1, pp. 181–186, 1972.

[19] X. Huang, M. Torgeir, and W. Cui, "An engineering model of fatigue crack growth under variable amplitude loading," *International Journal of Fatigue*, vol. 30, no. 1, pp. 2–10, 2008.

[20] S. Daneshpour, M. Koçak, S. Langlade, and M. Horstmann, "Effect of overload on fatigue crack retardation of aerospace Al-alloy laser welds using crack-tip plasticity analysis," *International Journal of Fatigue*, vol. 31, no. 10, pp. 1603–1612, 2009.

[21] B. K. C. Yuen and F. Taheri, "Proposed modifications to the Wheeler retardation model for multiple overloading fatigue life

prediction," *International Journal of Fatigue*, vol. 28, no. 10, pp. 1803–1819, 2006.

[22] F. J. McMaster and D. J. Smith, "Predictions of fatigue crack growth in aluminium alloy 2024-T351 using constraint factors," *International Journal of Fatigue*, vol. 23, no. 1, pp. S93–S101, 2001.

[23] J. Yang, W. Zhang, and Y. Liu, "Existence and insufficiency of the crack closure for fatigue crack growth analysis," *International Journal of Fatigue*, vol. 62, pp. 144–153, 2013.

[24] K. Sadananda, A. K. Vasudevan, R. L. Holtz, and E. U. Lee, "Analysis of overload effects and related phenomena," *International Journal of Fatigue*, vol. 21, no. 1, pp. S233–S246, 1999.

[25] Y. K. Tür and Ö. Vardar, "Periodic tensile overloads in 2024-T3 AL-alloy," *Engineering Fracture Mechanics*, vol. 53, no. 1, pp. 69–77, 1996.

[26] S. Daneshpour, J. Dyck, V. Ventzke, and N. Huber, "Crack retardation mechanism due to overload in base material and laser welds of Al alloys," *International Journal of Fatigue*, vol. 42, pp. 95–103, 2012.

[27] J. Schijve, M. Skorupa, A. Skorupa, T. Machniewicz, and P. Gruszczynski, "Fatigue crack growth in the aluminium alloy D16 under constant and variable amplitude loading," *International Journal of Fatigue*, vol. 26, no. 1, pp. 1–15, 2004.

[28] C. M. Manjunatha and B. K. Parida, "Prediction of fatigue crack growth after single overload in an aluminum alloy," *AIAA Journal*, vol. 42, no. 8, pp. 1536–1542, 2004.

[29] F. Taheri, D. Trask, and N. Pegg, "Experimental and analytical investigation of fatigue characteristics of 350WT steel under constant and variable amplitude loadings," *Marine Structures*, vol. 16, no. 1, pp. 69–91, 2003.

[30] P. A. Rushton and F. Taheri, "Prediction of crack growth in 350WT steel subjected to constant amplitude with over- and under-loads using a modified wheeler approach," *Marine Structures*, vol. 16, no. 7, pp. 517–539, 2003.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 142795, 15 pages
http://dx.doi.org/10.1155/2014/142795

Research Article

Intelligent Mechanical Fault Diagnosis Based on Multiwavelet Adaptive Threshold Denoising and MPSO

Hao Sun,[1,2] Ke Li,[1,2] Huaqing Wang,[3] Peng Chen,[4] and Yi Cao[1,2]

[1] *School of Mechanical Engineering, Jiangnan University, 1800 Li Hu Avenue, Wuxi, Jiangsu 214122, China*
[2] *Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Wuxi 214122, China*
[3] *School of Mechanical & Electrical Engineering, Beijing University of Chemical Technology, Chaoyang District, Beijing 100029, China*
[4] *Graduate School of Bioresources, Mie University, Mie 514-8507, Japan*

Correspondence should be addressed to Ke Li; dayanlv@live.cn and Huaqing Wang; wanghq_buct@hotmail.com

Received 16 April 2014; Revised 12 June 2014; Accepted 29 June 2014; Published 22 July 2014

Academic Editor: Weihua Li

The condition diagnosis of rotating machinery depends largely on the feature analysis of vibration signals measured for the condition diagnosis. However, the signals measured from rotating machinery usually are nonstationary and nonlinear and contain noise. The useful fault features are hidden in the heavy background noise. In this paper, a novel fault diagnosis method for rotating machinery based on multiwavelet adaptive threshold denoising and mutation particle swarm optimization (MPSO) is proposed. Geronimo, Hardin, and Massopust (GHM) multiwavelet is employed for extracting weak fault features under background noise, and the method of adaptively selecting appropriate threshold for multiwavelet with energy ratio of multiwavelet coefficient is presented. The six nondimensional symptom parameters (SPs) in the frequency domain are defined to reflect the features of the vibration signals measured in each state. Detection index (DI) using statistical theory has been also defined to evaluate the sensitiveness of SP for condition diagnosis. MPSO algorithm with adaptive inertia weight adjustment and particle mutation is proposed for condition identification. MPSO algorithm effectively solves local optimum and premature convergence problems of conventional particle swarm optimization (PSO) algorithm. It can provide a more accurate estimate on fault diagnosis. Practical examples of fault diagnosis for rolling element bearings are given to verify the effectiveness of the proposed method.

1. Introduction

Rolling element bearings are an important part of and widely used in rotating machinery. In practical application, bearing failures may cause the breakdown of equipment, and further, serious consequences may arise due to the failure. Thus, fault diagnosis and condition discrimination of bearings have an important significance for safe operation, guaranteeing production efficiency and reducing maintenance cost. Many reliability survey papers deal with failure statistics of rotating machinery subassemblies, focusing mainly on roller bearing because of their widespread use in industry [1–4]. Occurrence rate of bearing faults is very high in rotating machines, and other faults arising in rotation machines are often associated with bearing faults. In many instances, the accuracy of the instruments and devices used to monitor and control the rotation machines is highly dependent on the dynamic performance of bearings. Although fault diagnosis of rolling bearings is often artificially carried out using time or frequency analysis of vibration signals, there is a need for a reliable, fast automated diagnosis method.

Vibration diagnosis is commonly used to detect the faults and identify the states in rotating machine. The condition diagnosis of rotating machinery depends largely on the feature analysis of vibration signals measured for the condition diagnosis because the signals carry dynamic information about the machine state [5–7]. However, feature extraction for fault diagnosis is difficult, because if the vibration signals are measured at an early stage of the machine failure, or at a location away from the fault part, the vibration signals

contain strong noise. Stronger noise than the actual failure signal may lead to misrecognition of useful information for the condition diagnosis. Thus, it is important that the feature of the signal can be sensitively extracted at the state change of a machine.

Wavelet transform (WT) is well known for its ability to focus on localized structures in time-frequency domain which has been widely used for fault diagnosis of rolling element bearings [8–10]. It has the local characteristic of time domain as well as frequency domain and its time-frequency window is changeable. In the processing of nonstationary signals it presents better performance than the traditional Fourier analysis. However, the measured signals often contain strong noise and the fault features are hidden in the background noise, and it is not the best way for WT to match the different fault features with a single wavelet and scaling functions, which will reduce the fault diagnosis accuracy. Multiwavelet transform is the new development of WT. It is constructed from translations and dilations of scaling and wavelet vector functions and has the predominant properties such as orthogonality, symmetry, compact support, and higher order vanishing moments. Multiwavelet transform decomposes the signal into subsignals of different frequency bands based on vector basis functions, via inner product principle. Because of the multiple scaling and wavelet basis functions, multiwavelet transform has predominant advantages in feature extraction of signals. Recently multiwavelet transform has been applied in fault diagnosis of rotating machinery as a powerful tool. In [11], multiwavelet system was introduced to diagnose gear faults. In [12, 13], multiwavelet lifting scheme was improved for compound faults separation and extraction. In [14], the undecimated multiwavelet was proposed for fault diagnosis of planetary gearboxes.

PSO algorithm is a population based stochastic optimization technique developed by Kennedy and Eberhart in 1995 and inspired by social behavior of bird flocking or fish schooling [15]. In PSO algorithm, particles cooperate in finding good solutions for difficult discrete optimization problems. PSO algorithm has been applied to a variety of different problems, such as function optimization [16], scheduling [17], traveling salesman problem [18], neural network training [19, 20], and clustering task [21–23] which is the topic of interest in this paper. In recent years, PSO algorithm has been successfully applied in mechanical fault diagnosis; the domestic research on PSO fault diagnosis issues also has many articles reporting [24–27]. In [24], Bocaniala and Sa da Costa compared the time spent by PSO algorithm and genetic algorithm, testifying PSO algorithm with prominent superiority through fault diagnosis benchmark problem. In [25], Pan et al. used PSO algorithm to extract fault characteristics of rotation machinery. In [26, 27], PSO algorithm was used to diagnose gearbox fault. In this study, a clustering model is constructed by using an improved PSO called MPSO algorithm. It is used to classify the SPs calculated from the signals in each machine state for condition diagnosis, as well as obtaining their optimal clustering centers. According to these optimal clustering centers' information, the conditions of the machine can be accurately identified.

In order to extract the fault features of signals more effectively and identify mechanical condition more accurately, this paper proposes a novel fault diagnosis method for rotation machinery based on multiwavelet adaptive threshold denoising and MPSO algorithm. GHM multiwavelet is employed for extracting weak fault features under heavy background noise, and the method of adaptively selecting appropriate threshold values for multiwavelet with energy ratio of multiwavelet coefficient is presented. The six nondimensional SPs in the frequency domain are defined to reflect the features of the vibration signals measured in each state. DI using statistical theory has been also defined to evaluate the sensitiveness of SP for condition diagnosis. MPSO algorithm with adaptive inertia weight adjustment and particle mutation is proposed for condition identification. MPSO algorithm effectively solves local optimum and premature convergence problems of conventional particle swarm optimization (PSO) algorithm. It can provide a more accurate estimate on fault diagnosis. Practical examples of fault diagnosis for rolling element bearings are given to verify the effectiveness of the proposed method.

2. Feature Extraction by Multiwavelet Adaptive Threshold Denoising

2.1. Multiwavelet Theory. Multiwavelet consists of wavelet function vector Ψ and a function vector Φ is called multiscaling function. They are denoted as follows [28]:

$$\begin{aligned} \phi &= [\phi_1, \phi_2, \ldots, \phi_r]^T, \\ \psi &= [\psi_1, \psi_2, \ldots, \psi_r]^T. \end{aligned} \tag{1}$$

For a multiresolution of multiplicity $r > 1$.

Similar to scalar wavelet, Ψ and Φ satisfy the two-scale matrix refinement equations:

$$\begin{aligned} \phi(t) &= \sum_{k=0}^{N} H_K \phi(2t - k), \\ \psi(t) &= \sum_{k=0}^{N} G_K \phi(2t - k), \end{aligned} \tag{2}$$

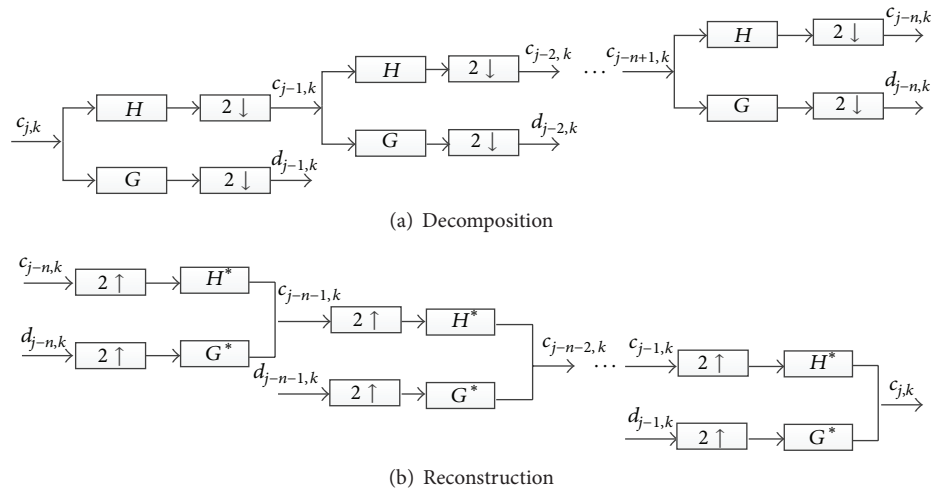
where $k \in Z$ and Z is the set of integers. H_K and G_K are low-pass and high-pass matrix filter banks, respectively.

Let $c_{j-1} = [c_{1,j-1}, \ldots, c_{r,j-1}]^T$ be the vector low frequency coefficients and let $d_{j-1} = [d_{1,j-1}, \ldots, d_{r,j-1}]^T$ be the vector high frequency coefficient; multiwavelet decomposition and composition are denoted as follows:

$$\begin{aligned} c_{j-1,n} &= \sum_k H_{k-2n} c_{j,k}, \qquad d_{j-1,n} = \sum_k G_{k-2n} c_{j,k}, \\ c_{j,k} &= \sum_n H^*_{k-2n} c_{j,k} + G^*_{k-2n} d_{j-1,n}, \end{aligned} \tag{3}$$

where $*$ means the complex conjugate transpose.

Figure 1 shows decomposition and reconstruction of multiwavelet.



(a) Decomposition

(b) Reconstruction

Figure 1: Decomposition and reconstruction of multiwavelet.

GHM multiwavelet constructed by Geronimo, Hardin, and Massopust is one of the most important multiwavelet systems with two pairs of scaling and wavelet functions and has the superior properties of short support, symmetry, orthogonality, and second approximation order [29]. Because of the excellent properties, GHM multiwavelets are adopted in this study. The multiscaling functions and multiwavelet functions of GHM multiwavelets are presented in Figure 2. The dilation and wavelet equations for GHM multiwavelet have four coefficients as follows:

$$\begin{bmatrix}\phi_1\\ \phi_2\end{bmatrix} = \sum_k H_k \begin{vmatrix}\phi_1(2t-k)\\ \phi_2(2t-k)\end{vmatrix},$$

$$H_0 = \begin{vmatrix}\frac{3}{10} & \frac{2\sqrt{2}}{5}\\ -\frac{\sqrt{2}}{40} & -\frac{3}{20}\end{vmatrix}, \qquad H_1 = \begin{vmatrix}\frac{3}{10} & 0\\ \frac{9\sqrt{2}}{40} & \frac{1}{2}\end{vmatrix},$$

$$H_2 = \begin{vmatrix}0 & 0\\ \frac{9\sqrt{2}}{40} & -\frac{3}{20}\end{vmatrix}, \qquad H_3 = \begin{vmatrix}0 & 0\\ -\frac{\sqrt{2}}{40} & 0\end{vmatrix},$$

$$\begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix} = \sum_k G_k \begin{vmatrix}\psi_1(2t-k)\\ \psi_2(2t-k)\end{vmatrix},$$

$$G_0 = \begin{vmatrix}-\frac{\sqrt{2}}{40} & -\frac{3}{20}\\ -\frac{1}{20} & -\frac{3\sqrt{2}}{20}\end{vmatrix}, \qquad G_1 = \begin{vmatrix}\frac{9\sqrt{2}}{40} & -\frac{1}{2}\\ \frac{9}{20} & 0\end{vmatrix},$$

$$G_2 = \begin{vmatrix}\frac{9\sqrt{2}}{40} & -\frac{3}{20}\\ -\frac{9}{20} & -\frac{3\sqrt{2}}{20}\end{vmatrix}, \qquad G_3 = \begin{vmatrix}-\frac{\sqrt{2}}{40} & 0\\ -\frac{1}{20} & 0\end{vmatrix}. \tag{4}$$

In view of the matrix filter banks, preprocessing is necessary to translate one stream input signal into two streams. Some preprocessing of preprocessing for multiwavelets has been proposed, such as repeated-row preprocessing, balanced multiwavelet, and prefilter methods [30]. Different preprocessing methods will produce different effect on performances of multiwavelets. It is a fundamental problem for each multiwavelet function to choose an appropriate preprocessing method for specific applications. In this study, the preprocessing method of repeated-row preferable for GHM multiwavelet is adopted and given as follows [31]:

$$s_n = \begin{vmatrix}x_n\\ cx_n\end{vmatrix}, \tag{5}$$

where x_n is original signal, s_n is the signal after preprocessing, and $c = \sqrt{2}$.

2.2. Multiwavelet Adaptive Threshold Denoising. Similar to single wavelet, multiwavelet denoising depends largely on the threshold denoising. The effect of threshold denoising depends on the selection of thresholds. A variety of threshold choosing methods can be mainly divided into two categories: global thresholding and level-dependent thresholding. The former chooses a single value of λ to be applied globally to all empirical wavelet coefficients, while the latter chooses different threshold value λ_t for each wavelet level. However, it is difficult to choose appropriate threshold values for different wavelet coefficient. A large threshold value cuts too many coefficients, resulting in the loss of useful information. Conversely, a too small threshold value will leave much noise. In this study, the method of adaptive selecting appropriate threshold values for multiwavelet denoising based on comparison of noise energy in different levels is proposed. According to the noise levels of wavelet coefficients, the adaptive threshold value is determined by energy ratio.

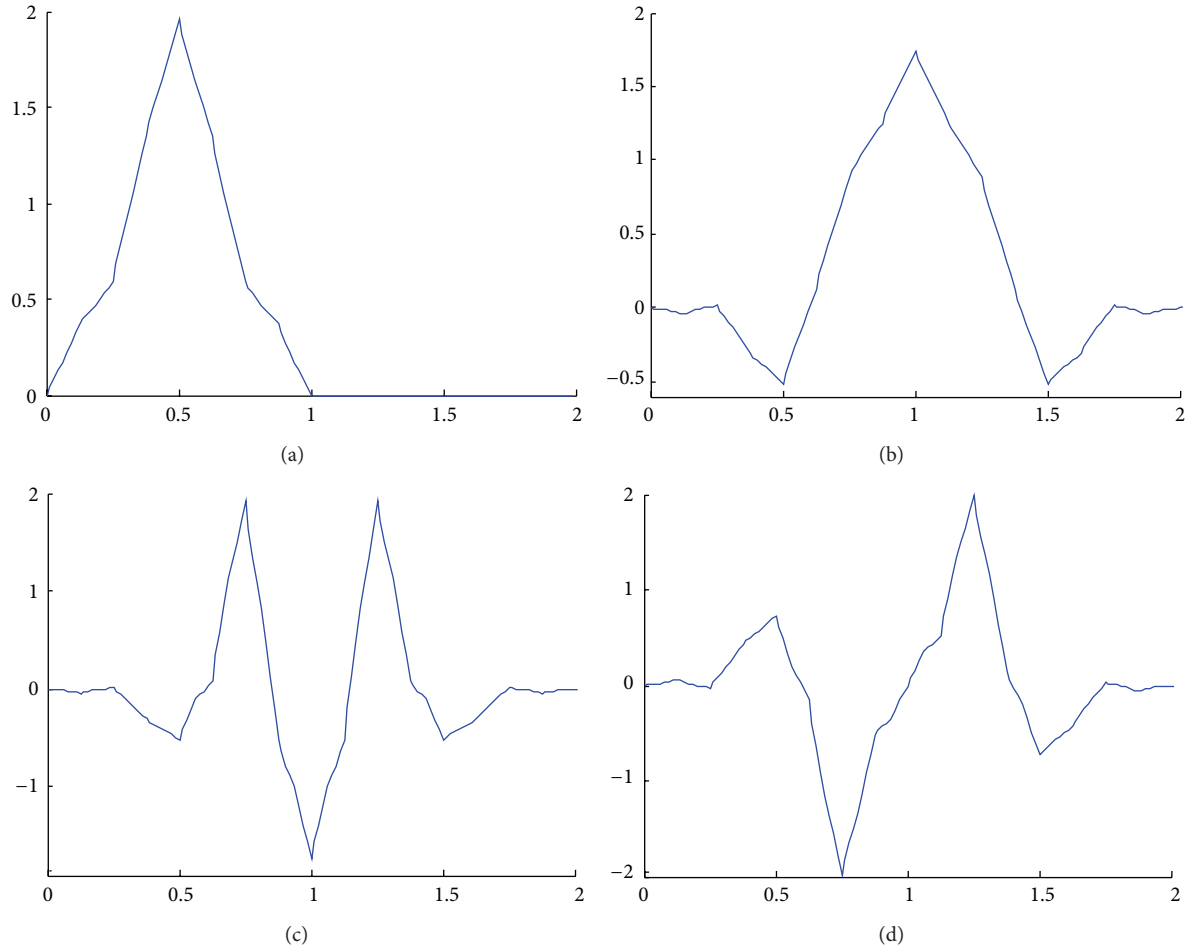


Figure 2: Multiscaling functions and multiwavelet functions of GHM. (a) Scaling function Φ_1, (b) scaling function Φ_2, (c) multiwavelet function Ψ_1, and (d) multiwavelet function Ψ_2.

The energy spectrum of multiwavelet coefficient is denoted as follows:

$$E^j = \sum_{i=1}^{r} E_i^j = \sum_{i=1}^{r} \sum_{n=1}^{m} |d_i(n)|^2, \tag{6}$$

where E^j represents multiwavelet coefficient total energy in the jth layer; E_i^j represents the multiwavelet coefficient energy of the i-dimensional in the jth layer; $d_j(n)$ is the multiwavelet coefficient in the jth layer after r-dimensional multiwavelet decomposition; and r is the number of dimensions of multiwavelet coefficient.

Energy ratio of multiwavelet coefficient can be obtained as follows:

$$p = \frac{E_i^j}{E^j}. \tag{7}$$

According to noise level of multiwavelet coefficients, the threshold values of each multiwavelet coefficient can be adaptively obtained as follows:

$$\mu = p \times \lambda_i^j,$$
$$\lambda_i^j = \frac{M \times \sqrt{2 \ln(n)}}{0.6745}, \tag{8}$$

where M is the median absolute value of multiwavelet coefficient; n is signal length.

In conclusion, the processing steps of multiwavelet adaptive threshold denoising are summarized as follows.

(1) Preprocess the original signal to transform it into two streams by the method of repeated-row preferable.

(2) Decompose two stream signals using multiwavelets.

(3) Threshold values are adaptively determined by energy ratio of the wavelet coefficients.

(4) Threshold the wavelet coefficients.

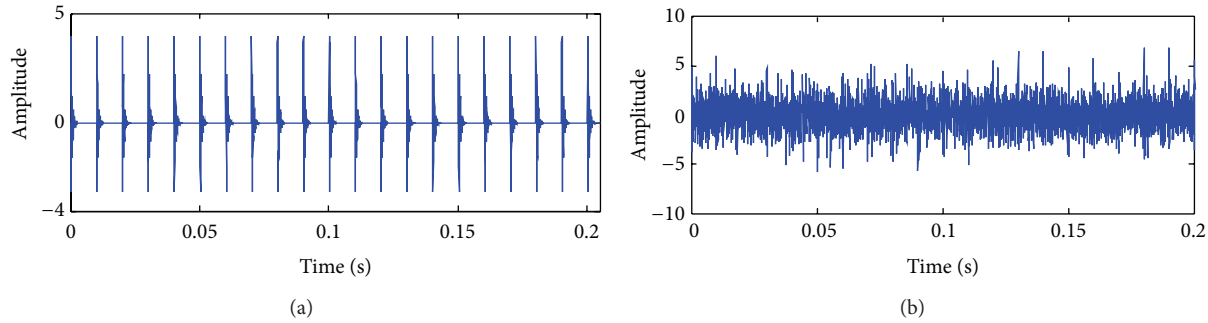


Figure 3: The simulation signal in the time domain: (a) the shock impulse signal; (b) the noisy signal.

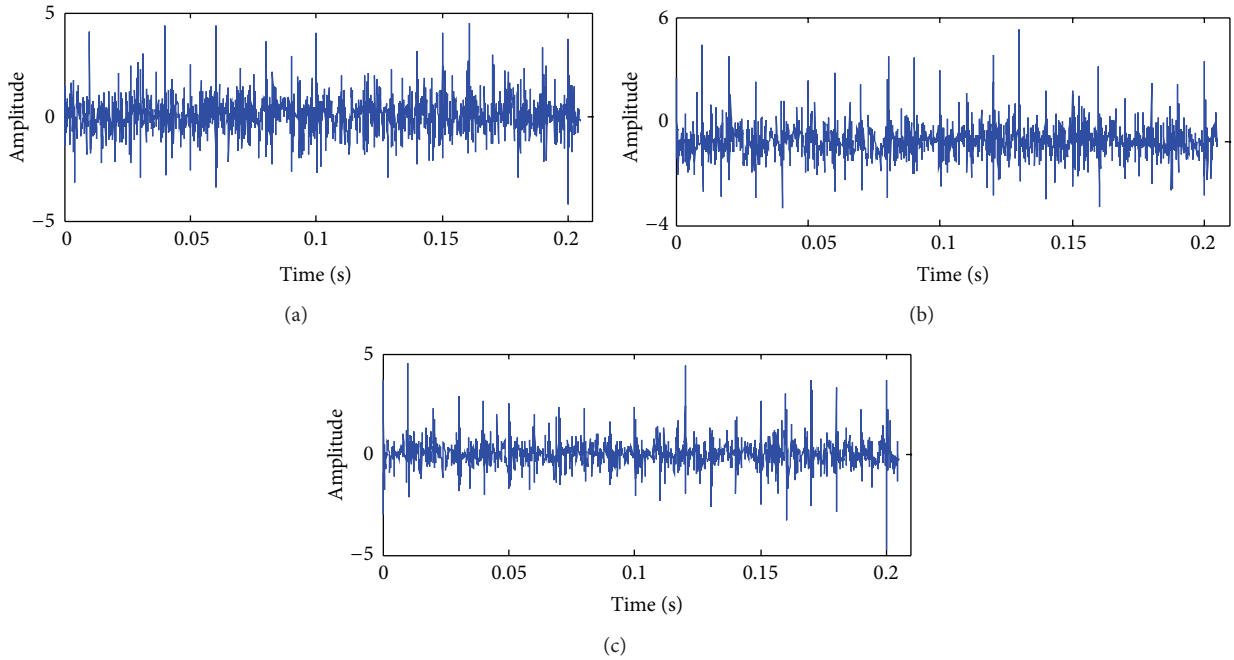


Figure 4: The denoising results using different wavelet denoising techniques. (a) Multiwavelet adaptive threshold denoising, (b) multiwavelet neighboring coefficient denoising, and (c) db2 wavelet threshold denoising.

(5) Reconstruct the threshold wavelet coefficients and the scale coefficients.

(6) Postprocess the two stream results to get the denoising signal.

In order to test effectiveness of multiwavelet adaptive threshold denoising proposed in this paper, a simulation experiment is designed as follows.

The simulation signal is composed of a periodic impulse component and white Gaussian noise to simulate a bearing fault. The periodic impulse signal with the period of 0.01 s is expressed as

$$x(t) = x_0 e^{-\xi\omega_n t} \sin \omega_n \sqrt{1-\xi^2}\, t, \tag{9}$$

where ξ is damp coefficient; ω_n denotes natural frequency; x_0 indicates displacement constant. The shock impulse signal is displayed in Figure 3(a). In this case, $\xi = 0.1$, $\omega_n = 3$ kHz, $x_0 = 5$, and sampling frequency and sampling points are 20 kHz and 4096, respectively. The simulation signal is shown in Figure 3(b), the signal has a low signal-noise ratio (SNR), and no useful features can be seen in the dynamic signal in the time domain.

The noisy signal is processed using GHM multiwavelet adaptive threshold denoising, GHM multiwavelet neighboring coefficient denoising, and Daubechies 2 (db2) wavelet threshold denoising, respectively. Type of thresholding used is soft thresholding, and decomposition level is four. Denoised signal's performance is evaluated based on mean square error (MSE) and SNR. Figure 4 shows the denoising results using different wavelet denoising techniques. The SNR and MSE of different wavelet denoising techniques are calculated, as shown in Table 1. The results indicate that the method of GHM multiwavelet adaptive threshold denoising has the maximum SNR and the minimum MSE, which means the method proposed in this study can effectively extract

Table 1: SNR and MSE of different wavelet denoising techniques.

	Multiwavelet adaptive threshold denoising	Multiwavelet neighboring coefficient denoising	db2 wavelet threshold denoising
SNR	14.516	11.098	9.667
MSE	0.212	0.235	0.306

the defect-induced shock impulses and eliminate much noise from the simulation signal.

3. Symptom Parameters for Fault Diagnosis and Sensitivity Evaluation

3.1. Symptom Parameters for Fault Diagnosis. When developing intelligent condition diagnosis system by computer, symptom parameters (SPs) are required to express the information indicated by a signal measured for diagnosing machinery faults. A good symptom parameter can correctly reflect states and the condition trend of plant machinery [32–34]. Many symptom parameters have been defined in the pattern recognition field. Here, six SPs in the frequency domain, commonly used for the fault diagnosis of plant machinery, are considered.

Frequency-domain skewness:

$$P_1 = \frac{\sum_{i=1}^{I} \left(f_i - \overline{f}\right)^3 \cdot F\left(f_i\right)}{\sigma^3 I}. \tag{10}$$

Frequency-domain kurtosis:

$$P_2 = \frac{\sum_{i=1}^{I} \left(f_i - \overline{f}\right)^4 \cdot F\left(f_i\right)}{\sigma^4 \cdot I}. \tag{11}$$

Mean frequency that wave shape cross the mean of time-domain signal:

$$P_3 = \sqrt{\frac{\sum_{i=1}^{I} f_i^4 \cdot F\left(f_i\right)}{\sum_{i=1}^{I} f_i^2 \cdot F\left(f_i\right)}}. \tag{12}$$

Stabilization factor of wave shape:

$$P_4 = \frac{\sum_{i=1}^{I} f_i^2 \cdot F\left(f_i\right)}{\sqrt{\sum_{i=1}^{I} F\left(f_i\right) \sum_{i=1}^{I} f_i^4 \cdot F\left(f_i\right)}}. \tag{13}$$

Sum of the squares of the power spectrum:

$$P_5 = \sum_{i=1}^{I} F\left(f_i\right). \tag{14}$$

Square root of the sum of the squares of the power spectrum:

$$P_6 = \sqrt{\sum_{i=1}^{I} F^2\left(f_i\right)}, \tag{15}$$

where I is the number of spectrum lines, f_i is frequency, and from 0 Hz to the maximum analysis frequency, $F(f_i)$ is the power spectrum value at frequency f_i, and $i = 1 \sim I$. $\overline{f}$ is mean value of the analysis frequency, and $\overline{f} = (\sum_{i=1}^{I} f_i \cdot F(f_i)) / \sum_{i=1}^{I} F(f_i)$; σ is standard deviation, and $\sigma = \sqrt{(\sum_{i=1}^{I} (f_i - \overline{f})^2 \cdot F(f_i))/I}$.

3.2. Detection Index. For automatic diagnosis, SPs are needed that can sensitively distinguish the fault types. In order to evaluate the sensitivity of a SP for distinguishing two states, such as a normal or an abnormal state, DI is defined as follows.

Supposing that x_1 and x_2 are the SP values calculated from the signals measured in state 1 and state 2, respectively, their average value and standard deviation are μ and σ. The DI is calculated by

$$\mathrm{DI} = \frac{\left|\mu_1 - \mu_2\right|}{\sqrt{\sigma_1 + \sigma_2}}. \tag{16}$$

The distinction rate (DR) is defined as

$$\mathrm{DR} = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\mathrm{DI}} \exp\left(-\frac{\mu^2}{2}\right) d_\mu. \tag{17}$$

It is obvious that the larger the value of the DI, the larger the value of the DR will be and, therefore, the better the SP will be. Thus, the DIcan be used as the index of the quality to evaluate the distinguishing sensitivity of the SP.

The number of symptom parameters used for diagnosis and fault types are M and N, respectively; the synthetic detection index (SDI) is defined as follows:

$$\mathrm{SDI} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{k=1}^{M} \frac{\left|\mu_{ik} - \mu_{jk}\right|}{\sqrt{\sigma_{ik}^2 + \sigma_{jk}^2}}. \tag{18}$$

4. MPSO for Condition Diagnosis

4.1. Brief of PSO. PSO algorithm is based on the groups, and according to the environmental fitness, individual in groups will be moved to the good region. The algorithm evaluates the optimal result by using evolutionary fitness function of group, and each particle in the algorithm has a fitness value determined by the fitness function; two properties of position and speed that are used to show the position and moving speed of the current articles in the solving space, by the fitness function value corresponding to particle position coordinate, determine the performance of particles. In PSO algorithm, each particle adjusts its position according

to its own experience and according to the experience of a neighboring particle, making use of the best position encountered by itself and its neighbor.

In the R-dimensional search space, the i particle's space position is defined as follows:

$$P(i) \cdot \text{location} = \left[X_{i1}, X_{i2}, \ldots, X_{iR}\right]. \tag{19}$$

The velocity of particle i is defined as follows:

$$P(i) \cdot \text{velocity} = \left[V_{i1}, V_{i2}, \ldots, V_{iR}\right]. \tag{20}$$

The best previous position of particle i is defined as follows:

$$P(i) \cdot \text{best} = \left[P_{i1}, P_{i2}, \ldots, P_{iR}\right]. \tag{21}$$

The best position among all particles experienced is defined as follows:

$$g(i) \cdot \text{best} = \left[g_{i1}, g_{i2}, \ldots, g_{iR}\right]. \tag{22}$$

The particle updates the position and velocity according to the following equations:

$$\begin{aligned} P(i) &\cdot \text{velocity}(t+1) \\ &= \omega P(i) \cdot \text{velocity}(t) \\ &\quad + \eta_1 r_1 \left[P(i) \cdot \text{best}(t) - P(i) \cdot \text{location}(t)\right] \\ &\quad + \eta_2 r_2 \left[g(i)\,\text{best}(t) - P(i) \cdot \text{location}(t)\right], \\ P(i) &\cdot \text{location}(t+1) \\ &= P(i) \cdot \text{location}(t) + P(i) \cdot \text{velocity}(t+1), \end{aligned} \tag{23}$$

where r_1 and r_2 are the random numbers within $(0, 1)$ and η_1 and η_2 are the acceleration which constants the control of how far a particle moves in a single generation. The inertia weight ω controls the previous velocity of particle, and it is defined as follows:

$$\omega = 0.5 + \frac{\text{rand}}{2}, \tag{24}$$

where rand is random generated number between 0 and 1.

Although PSO algorithm is easy to realize, the method is easy to trap into local optimum. Shi and Eberhart proposed a linearly decreasing weight particle swarm optimization (WPSO) of which a linearly decreasing inertia factor was introduced into the velocity of the updated equation from the original PSO [35, 36]. The performance of WPSO is significantly improved over the original PSO because WPSO balances out the global and local search abilities of the swarm effectively. The equation for the linearly decreased weight is defined as follows:

$$\omega_l = \omega_{\max} - \text{iteration} \times \frac{\omega_{\max} - \omega_{\min}}{\text{iteration}_{\max}}, \tag{25}$$

where $\omega_{\max}$ is 1, $\omega_{\min}$ is 0.1, and $\text{iteration}_{\max}$ is the maximum number of the allowed iterations.

The velocity of the updated equation for WPSO is defined as follows:

$$\begin{aligned} P(i) &\cdot \text{velocity}(t+1) \\ &= \omega_l P(i) \cdot \text{velocity}(t) \\ &\quad + \eta_1 r_1 \left[P(i) \cdot \text{best}(t) - P(i) \cdot \text{location}(t)\right] \\ &\quad + \eta_2 r_2 \left[g(i)\,\text{best}(t) - P(i) \cdot \text{location}(t)\right]. \end{aligned} \tag{26}$$

4.2. MPSO. Although WPSO algorithm improved conventional PSO to a certain extent, it cannot adapt to all of complex practical problems. The main reasons can be explained as follows. (1) The inertia weight of conventional WPSO algorithm is monotone decreasing, and adjustment ability of WPSO algorithm is limited. If particles cannot find optimal point in the initial stage of the algorithm, WPSO algorithm is easy to trap into local optimum with the decrease of the inertia weight. (2) With increasing iterations, particle diversity of WPSO algorithm decreases; it causes deterioration of global search ability; WPSO algorithm is also easy to trap into local optimum and premature convergence.

To improve global search ability and adjustment ability of conventional PSO algorithm and prevent local optimum and premature convergence problems, MPSO algorithm with adaptive inertia weight adjustment and particle mutation is proposed in this paper.

4.2.1. Adaptive Inertia Weight. Define change rate of fitness value:

$$R = \frac{\left|f(t+5) - f(t)\right|}{\left|f(t)\right|}, \tag{27}$$

where $f(t)$ is optimum fitness value of the tth iteration; $f(t+5)$ is optimum fitness value of the $(t+5)$th iteration; R indicates change rate of fitness value in five iterations.

According to the variation of R, the inertia weight ω adaptively adjusts as follows:

$$\omega = \begin{cases} k_1 + 0.5q, & R > 0.05, \\ k_2 + 0.5q, & R \le 0.05, \end{cases} \tag{28}$$

where q is a random number with a uniform probability within 0~1; k_1 and k_2 are parameters; $k_1 > k_2$; the choice of k_1 and k_2 is determined experimentally; here $k_1 = 0.5$ and $k_2 = 0.2$. When $R > 0.05$, the algorithm is in the exploration stage, and a large ω is beneficial to the algorithm's convergence. When $R \le 0.05$, the algorithm is in the development stage, and a small ω is beneficial to searching optimum point.

4.2.2. Particle Mutation. To increase particle diversity of PSO algorithm, the method of particle mutation is proposed. In the operation process of PSO algorithm, if the best position among all particles $g$best does not change in a long time, some particles are mutated according to a certain probability. The execution process of the mutation for PSO is as follows.

(1) All particles are arranged in ascending order according to the values of the fitness function.

Table 2: DIs of each SP.

	P_1	P_2	P_3	P_4	P_5	P_6
$DI_{N:O}$	**5.733**	2.607	0.953	13.973	**6.920**	**3.287**
$DI_{N:I}$	**1.947**	1.467	0.740	1.593	**3.387**	**1.840**
$DI_{N:R}$	**4.540**	3.580	0.513	0.707	**2.287**	**2.820**
$DI_{O:I}$	**3.793**	0.467	0.587	1.367	**2.413**	**1.680**
$DI_{O:R}$	**2.007**	0.693	1.040	1.533	**1.567**	**1.606**
$DI_{I:R}$	**1.560**	0.467	0.813	1.087	**1.687**	**1.413**

(2) The m ($m > 1$) particles with smaller fitness functions are selected.

(3) Random data r_i $\{i = 1, 2 \ldots, m\}$ for selected particles are produced automatically.

(4) A weight P_m is set, and $0.1 < P < 0.5$.

(5) P_m is compared with r_i, if $P_m > r_i$, and then the particle's space position is updated by using (29).

(6) Steps (3)–(5) are looped until the space position of m particles are updated:

$$x_{ij}^{t+1} = x_{ij}^{t}\left(1 + 0.5\eta\right), \tag{29}$$

where η is random data that obeys Gaussian$(0, 1)$ distribution.

4.3. Fitness Function of MPSO for Condition Diagnosis. Assume that N is the sample set of vibration signals measured in m different states; the length of N is n, $N = \{s_1, s_2, \ldots, s_n\}$. Every sample signal has t identified symptoms (in this paper, the symptoms are P_1–P_6). Then, the clustering analysis is to divide n sample data into m states, such that the fitness function F shown in (30) is minimized:

$$\min F = \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{t} a_{ij}\left\|S_{ik} - X_{jk}\right\|^2, \tag{30}$$

$$X_{jk} = \frac{\sum_{i=1}^{n} a_{ij}S_{ik}}{\sum_{i=1}^{n} a_{ij}} \quad \left(j = 1, 2, \ldots, m;\ k = 1, 2, \ldots, t\right), \tag{31}$$

$$a_{ij} = \begin{cases} 1, & \text{if } S_i \in \text{state } j \\ 0, & \text{if } S_i \notin \text{state } j \end{cases} \quad \left(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m\right). \tag{32}$$

5. Diagnosis and Application

In this section, the application of condition diagnosis for a rolling bearing is shown to verify that the method proposed in this paper is effective.

5.1. Experimental System. Figure 5 shows the experimental system for a roller bearing fault diagnosis test. The most commonly occurring faults in a roller element bearing are the outer-race defect, the inner-race defect, and the roller element defect. These fault bearings are shown in Figure 6 and were created artificially using a wire-cutting machine. In this work

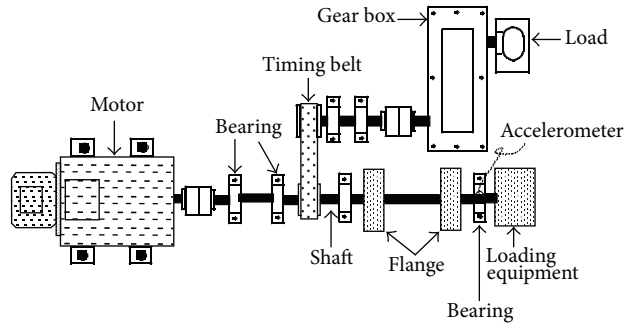


Figure 5: Experimental system for bearing fault diagnosis.

an accelerometer (PCB MA352A60) was used to measure the vibration signals of the vertical direction in the normal, the outer-race defect, the inner-race defect, and the roller element defect states, respectively. The original vibration signals in each state are measured at a constant speed (800 rpm), and a 150 kg load was also transported on the rotating shaft by the loading equipment (RCS2-RA13R) while the vibration signals were being measured. The sampling frequency of the signal measurement was 50 kHz, and the sampling time was 20 s. All of the data was divided to 100 parts; 40 parts were used to train diagnosis system; other parts were used for condition identification test. Spectrum values at frequency Figures 7(a), 8(a), 9(a), and 10(a) show the original vibration signal in each state, and Figures 7(b), 8(b), 9(b), and 10(b) show the multiwavelet adaptive threshold denoising results of the vibration signal in each state.

5.2. Diagnosis by the Proposed Method. The main procedure for fault diagnosis using GHM multiwavelet adaptive threshold denoising and MPSO algorithm is shown in Figure 11 and explained as follows.

(1) Vibration signals are measured in each known state.

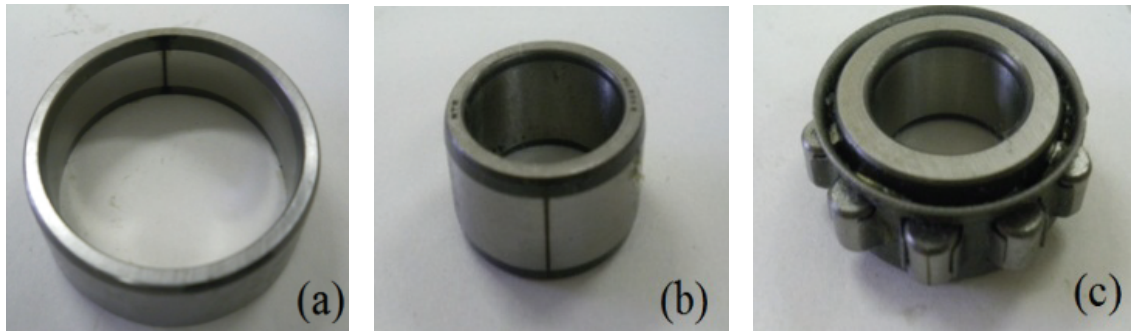
(2) Weak fault feature is extracted by using GHM multiwavelet adaptive threshold denoising.

(3) SPs are calculated using (10)–(15).

(4) The highly sensitive SPs are selected for condition diagnosis by DI.

(5) MPSO algorithm is trained with SPs selected by DI, and the optimal clustering centers are obtained.

(6) Condition of the bearing can be diagnosed by the trained MPSO algorithm and SPs.



(a) (b) (c)

Figure 6: Bearing defects. (a) Outer-race defect. (b) Inner-race defect. (c) Roller defect.

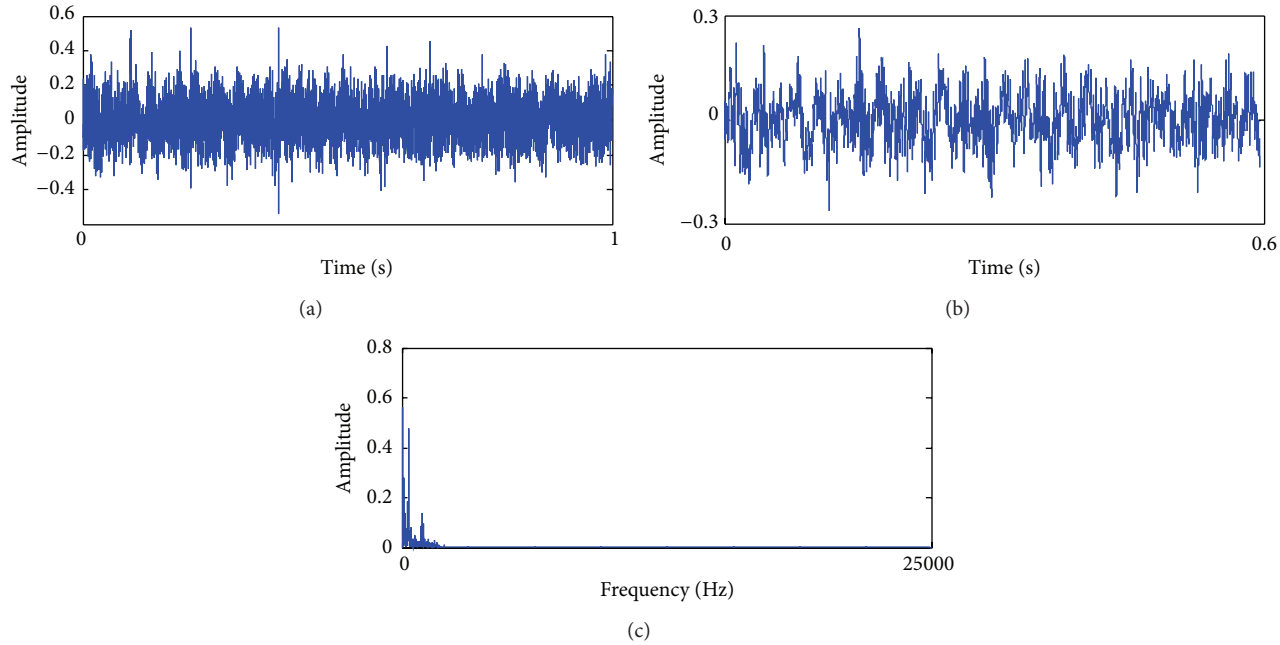


Figure 7: The vibration signal in normal state: (a) the original vibration signal, (b) after multiwavelet adaptive threshold denoising, and (c) Fourier spectrum of the denoised signal.

In this study, the good SPs which have high sensitivity for distinguishing each fault state of the bearing are selected by the method of DI. As an example, Table 2 lists parts of DIs of SPs. The maximum value (50.49) of SDI is obtained in the case of the combination of P_1, P_5, and P_6, and when P_1, P_5, and P_6, are used for distinguishing each state separately, the DIs are larger than 1.41, and all of the DRs are larger than 92.1%. Therefore, P_1, P_5, and P_6 have high sensitivity for distinguishing each fault state of the bearing.

In this study, MPSO automatically obtains the optimal clustering centers according to the classification of the sample data information. The purpose of training MPSO is the acquisition of optimum clustering centers. The SPs selected by DI were input into MPSO. MPSO converged to the optimum clustering centers. In the training process of MPSO, at first, the sample data are classified into the normal, the outer-race defect, the inner-race defect, and the roller element defect randomly. The fitness values and the clustering centers are calculated by (30) and (31). With increasing iterations, the speed and position (classification of the sample data) of the particle are updated incessantly, and according to the classification of the sample data information, the clustering centers are also updated. Finally, the optimal clustering centers with a minimum fitness value are calculated.

To explain the effectiveness of MPSO algorithm, a comparison is made among MPSO, WPSO, and PSO algorithms. The optimal clustering centers of each state are obtained by MPSO, WPSO, and PSO algorithms, respectively. Particle number and iteration number of the three algorithms are 50 and 1000, respectively. The clustering centers obtained by each method are shown in Tables 3, 4, and 5. Figure 12 shows the fitness curve of PSO, WPSO, and MPSO algorithms. It is obvious that MPSO algorithm has the minimum fitness value; namely, the optimal clustering centers obtained by

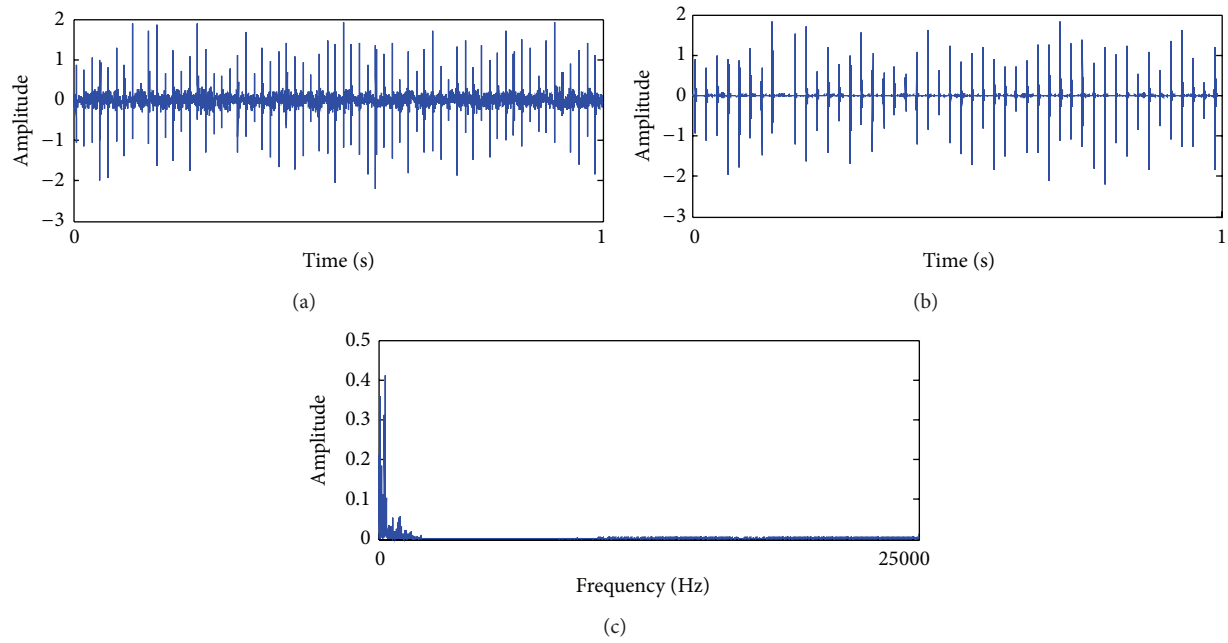


Figure 8: The vibration signal in outer-race defect state: (a) original vibration signal, (b) after multiwavelet adaptive threshold denoising, and (c) Fourier spectrum of the denoised signal.

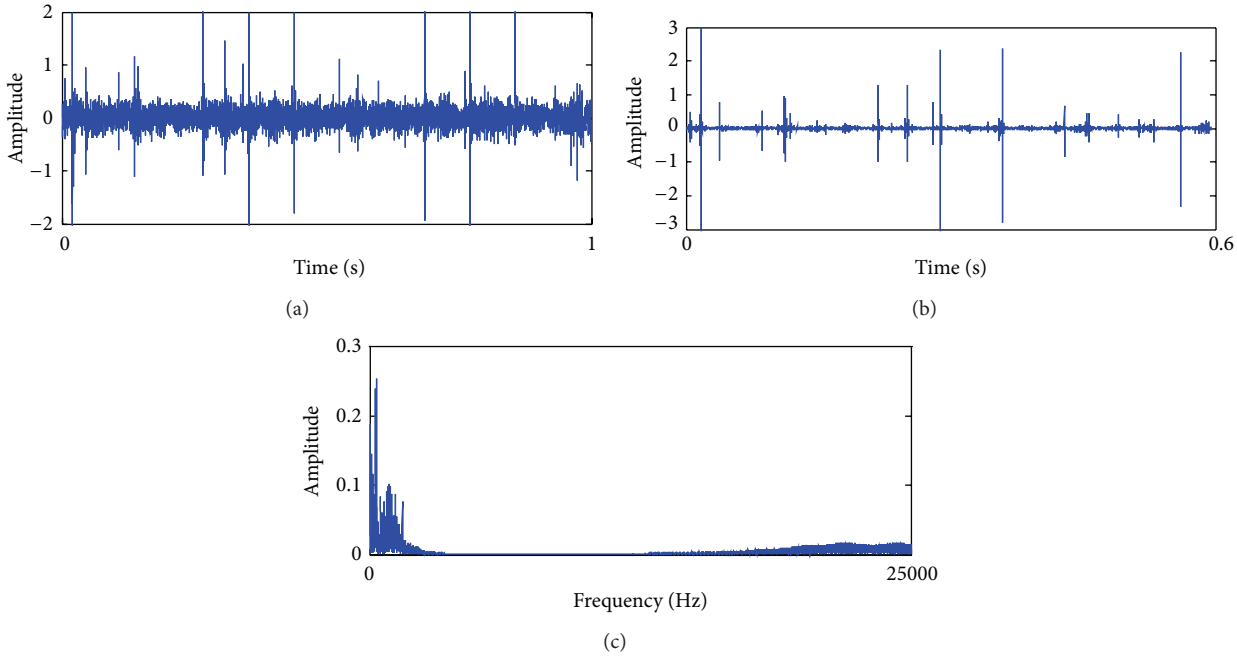


Figure 9: The vibration signal in inner-race defect state: (a) original vibration signal, (b) after multiwavelet adaptive threshold denoising, and (c) Fourier spectrum of the denoised signal.

MPSO algorithm are the most accurate. MPSO algorithm has stronger capacity of searching optimal solution.

After training MPSO algorithm, to verify the diagnostic capability of the proposed method in this paper, the test data measured in each known state that had not been used to train MPSO algorithm were used. When inputting the test data into the trained MPSO algorithm, MPSO classified the test data according to the information of the optimum clustering centers shown in Table 3 and correctly and quickly output identification results. As an example, some diagnosis results are listed in Table 6. We also identified condition of the rolling bearing using PSO and WPSO algorithms, respectively, and

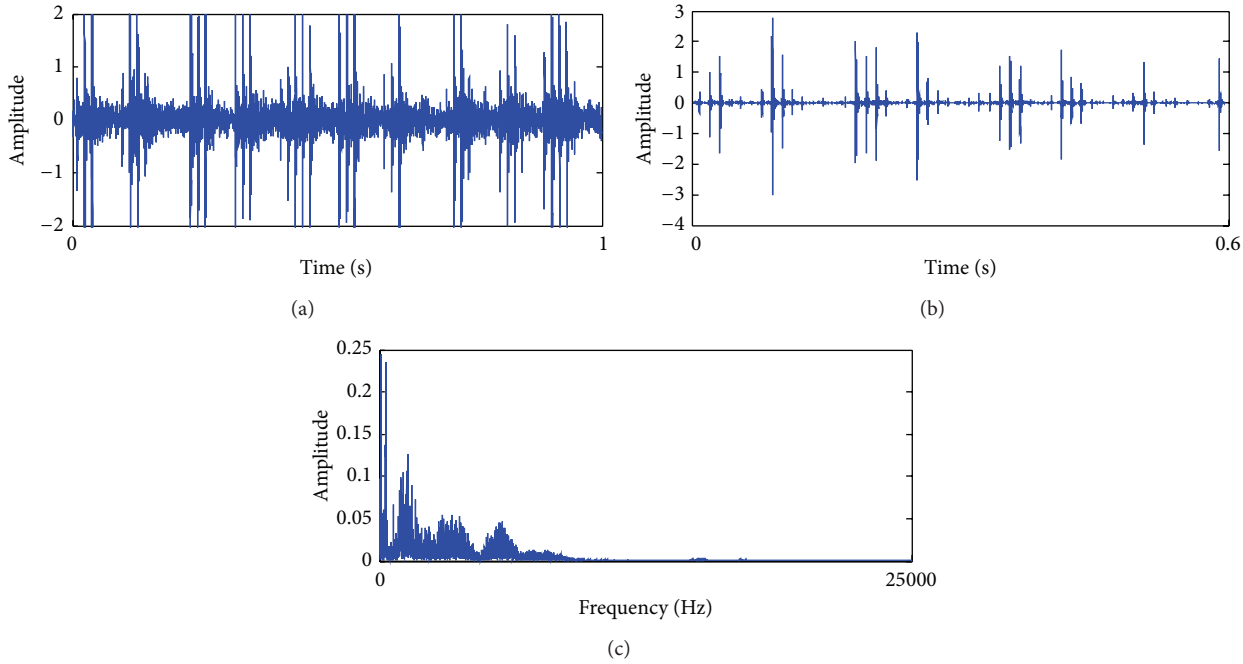


Figure 10: The vibration signal in roller defect state: (a) original vibration signal, (b) after multiwavelet adaptive threshold denoising, and (c) Fourier spectrum of the denoised signal.

Table 3: Clustering centers obtained by MSPO.

Machinery condition	Clustering centers		
	P_1	P_5	P_6
Normal	0.378	2.882	0.076
Outer-race defect	0.433	6.094	0.571
Inner-race defect	0.118	4.051	0.187
Roller element defect	0.729	5.246	0.022

Table 4: Clustering centers obtained by SPO.

Machinery condition	Clustering centers		
	P_1	P_5	P_6
Normal	0.371	2.922	0.069
Outer-race defect	0.583	6.081	0.558
Inner-race defect	0.082	3.825	0.175
Roller element defect	0.579	5.325	0.020

Table 5: Clustering centers obtained by WSPO.

Machinery Condition	Clustering Centers		
	P_1	P_5	P_6
Normal	0.395	2.831	0.073
Outer-race defect	0.558	6.105	0.583
Inner-race defect	0.155	4.228	0.223
Roller element defect	0.796	5.045	0.021

some identification results are shown in Tables 7 and 8. The comparison of diagnostic capability of each method is shown in Figure 13. Viewing the overall diagnostic results, diagnostic accuracy of each state using MPSO algorithm is 100%, 88.3%, 86.7%, and 81.7%, respectively; they are the largest in three methods. The method proposed in this study provides a more accurate estimate in the case of the rolling bearing faults diagnosis. These results verified the efficiency of the intelligent diagnosis method using multiwavelet adaptive threshold denoising and MPSO proposed in this paper.

6. Conclusions

In order to diagnose faults of rotation machinery at an early stage, this paper proposed a novel intelligent condition diagnosis method using multiwavelet adaptive threshold denoising and MPSO to detect faults and distinguish fault types at an early stage. The main conclusions are summarized as follows.

(1) The method of multiwavelet adaptive threshold denoising was presented for extracting weak fault features under background noise. It could adaptively select appropriate threshold for multiwavelet with energy ratio of multiwavelet coefficient. The simulation experiment verified that the method of multiwavelet adaptive threshold denoising can effectively extract fault features and eliminate much noise from the noisy signal.

(2) The six SPs in the frequency domain were defined for reflecting the features of vibration signals measured in each state. DI using statistical theory had been also defined to evaluate the applicability of the SPs for the condition diagnosis measured in each state. DI could

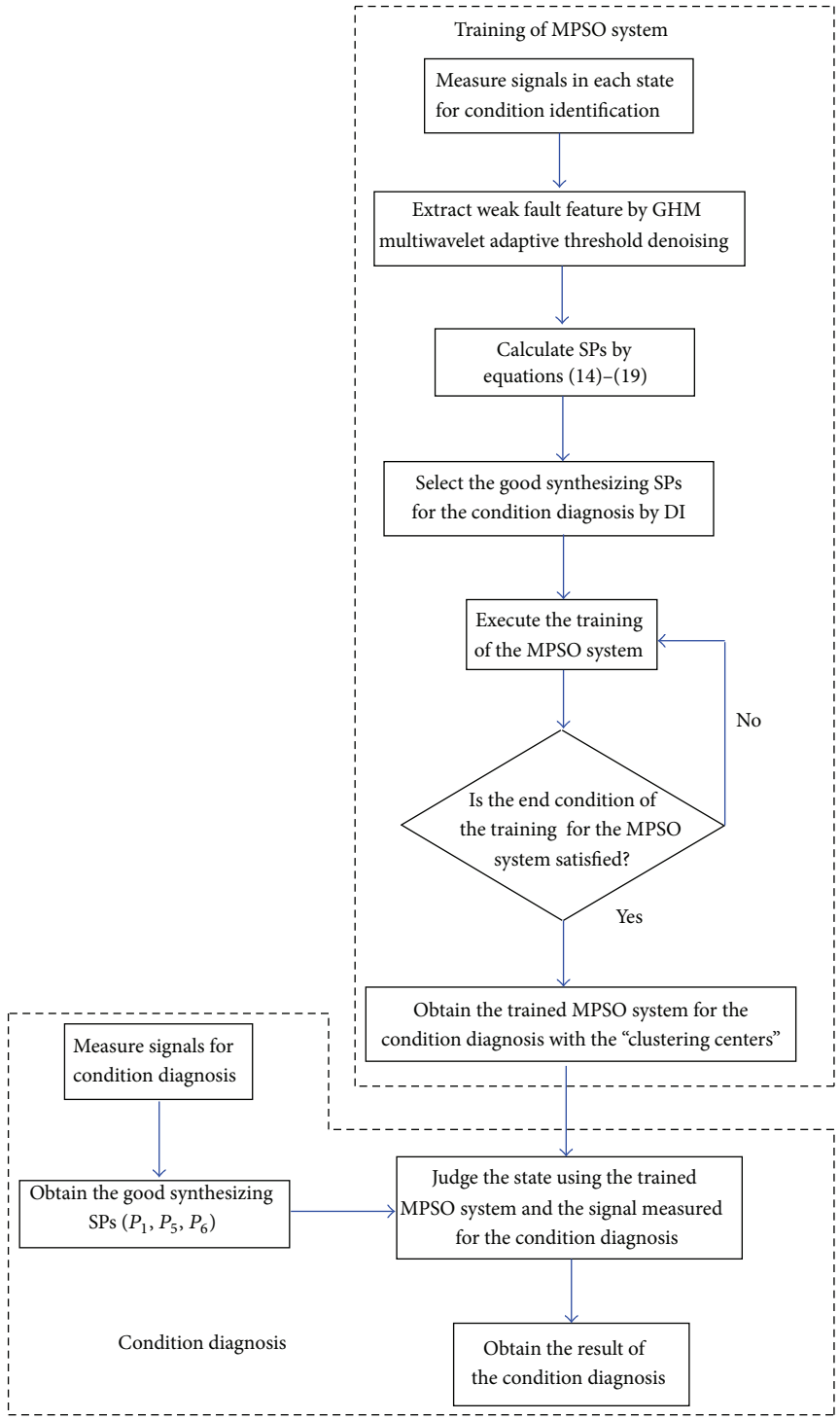


Figure 11: Flowchart for the condition diagnosis by the method proposed in this study.

be used to indicate the fitness of a SP for condition identification.

(3) MPSO algorithm with adaptive inertia weight adjustment and particle mutation was proposed for condition identification. MPSO algorithm was used to classify the SPs calculated from the signals in each machine state for condition diagnosis, as well as obtaining their optimal clustering centers. According to these optimal clustering centers' information, the conditions of rotation machinery could be accurately identified. MPSO algorithm effectively solved local

Table 6: Diagnosis result using proposed method.

Machinery condition	Number of training data	Number of test data	Number of correct results	Diagnostic accuracy (%)
Normal	40	60	60	100%
Outer-race defect	40	60	53	88.3%
Inner-race defect	40	60	52	86.7%
Roller element defect	40	60	49	81.7%

Table 7: Diagnostic result using PSO.

Machinery condition	Number of training data	Number of test data	Number of correct results	Diagnostic accuracy (%)
Normal	40	60	52	86.7%
Outer-race defect	40	60	43	71.6%
Inner-race defect	40	60	41	68.3%
Roller element defect	40	60	43	71.6%

Table 8: Diagnosis result using WPSO.

Machinery condition	Number of training data	Number of test data	Number of correct results	Diagnostic accuracy (%)
Normal	40	60	60	100%
Outer-race defect	40	60	49	81.7%
Inner-race defect	40	60	45	75%
Roller element defect	40	60	46	76.7%

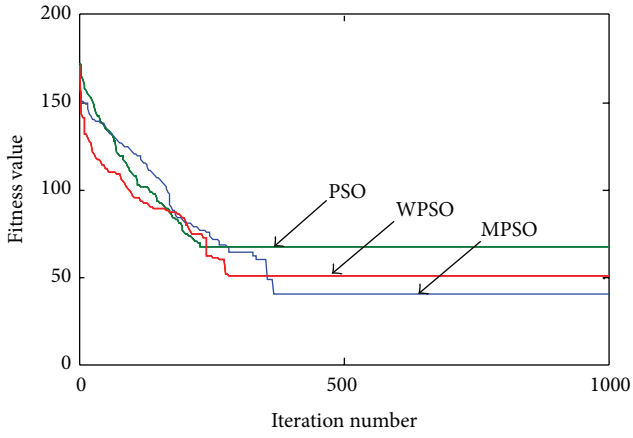


Figure 12: The fitness curve of PSO, WPSO, and MPSO algorithms.

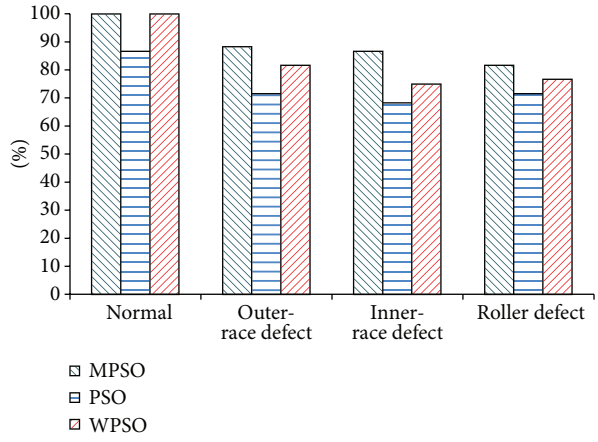


Figure 13: Comparison of diagnostic capability.

optimum and premature convergence problems of conventional PSO algorithm and raised diagnostic accuracy.

(4) Practical example of condition diagnosis for a rolling bearing verified that the method proposed in this paper was effective. Moreover, a comparison was also made among MPSO, WPSO, and PSO algorithms. The diagnostic results show that MPSO algorithm could provide a more accurate estimate in the case of the rolling bearing faults diagnosis.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work is supported by Fundamental Research Funds for the Central Universities (Grant nos. JUSRP51316B and JUSesRP1056).

References

[1] M. Pacas, S. Villwock, and R. Dietrich, "Bearing damage detection in permanent magnet synchronous machines," in *Proceedings of the IEEE Energy Conversion Congress and Exposition (ECCE '09)*, pp. 1098–1103, September 2009.

[2] J. R. Stack, T. G. Habetler, and R. G. Harley, "Fault-signature modeling and detection of inner-race bearing faults," *IEEE Transactions on Industry Applications*, vol. 42, no. 1, pp. 61–68, 2006.

[3] K. Li, P. Chen, S. Wang, and H. Wang, "Intelligent diagnosis method for bearing using non-dimensional symptom parameters and Ant Colony Optimization," *Information*, vol. 15, no. 2, pp. 867–877, 2012.

[4] C. Bianchini, F. Immovilli, M. Cocconcelli, R. Rubini, and A. Bellini, "Fault detection of linear bearings in brushless AC linear motors by vibration analysis," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 5, pp. 1684–1694, 2011.

[5] B. Liu and S.-F. Ling, "On the selection of informative wavelets for machinery diagnosis," *Mechanical Systems and Signal Processing*, vol. 13, no. 1, pp. 145–162, 1999.

[6] J. Lin and L. Qu, "Feature extraction based on morlet wavelet and its application for mechanical fault diagnosis," *Journal of Sound and Vibration*, vol. 234, no. 1, pp. 135–148, 2000.

[7] Q. B. Zhu, "Gear fault diagnosis system based on wavelet neural networks," *Dynamics of Continuous Discrete and Impulsive Systems-series A-Mathematical Analysis*, vol. 13, part 2, pp. 671–673, 2006.

[8] K. Li, P. Chen, and H. Wang, "Intelligent diagnosis method for rotating machinery using wavelet transform and ant colony optimization," *IEEE Sensors Journal*, vol. 12, no. 7, pp. 2474–2484, 2012.

[9] C. Junsheng, Y. Dejie, and Y. Yu, "Application of an impulse response wavelet to fault diagnosis of rolling bearings," *Mechanical Systems and Signal Processing*, vol. 21, no. 2, pp. 920–929, 2007.

[10] X. Fan, M. Liang, T. H. Yeap, and B. Kind, "A joint wavelet lifting and independent component analysis approach to fault detection of rolling element bearings," *Smart Materials and Structures*, vol. 16, no. 5, pp. 1973–1987, 2007.

[11] S. E. Khadem and M. Rezaee, "Development of vibration signature analysis using multiwavelet systems," *Journal of Sound and Vibration*, vol. 261, no. 4, pp. 613–633, 2003.

[12] J. Yuan, Z. J. He, and Y. Y. Zi, "Separation and extraction of electromechanical equipment compound faults using lifting multiwavelets," *Journal of Mechanical Engineering*, vol. 46, no. 1, pp. 79–85, 2010.

[13] H. Jiang, C. Li, and H. Li, "An improved EEMD with multiwavelet packet for rotating machinery multi-fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 36, no. 2, pp. 225–239, 2013.

[14] H. Sun, Y. Zi, J. Yuan, Z. He, K. Li, and X. Chen, "Undecimated multiwavelet and Hilbert-Huang time-frequency analysis and its application in the incipient fault diagnosis of planetary gearboxes," *Journal of Mechanical Engineering*, vol. 49, no. 3, pp. 56–62, 2013.

[15] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1942–1948, December 1995.

[16] P. Shao and Z. J. Wu, "Rosenbrock function optimization based on improved particle swarm optimization algorithm," *Computer Science*, vol. 40, no. 9, pp. 194–197, 2013.

[17] C. Akjiratikarl, P. Yenradee, and P. R. Drake, "PSO-based algorithm for home care worker scheduling in the UK," *Computers and Industrial Engineering*, vol. 53, no. 4, pp. 559–583, 2007.

[18] W. Pang, K. Wang, C. Zhou, and L. Dong, "Fuzzy discrete particle swarm optimization for solving traveling salesman problem," in *Proceedings of the 4th International Conference on Computer and Information Technology (CIT '04)*, pp. 796–800, September 2004.

[19] M. Ma and L. Zhang, "Particle swarm optimization algorithm design for fuzzy neural network," *Fuzzy Information and Engineering Advances in Soft Computing*, vol. 40, pp. 309–314, 2007.

[20] Q. Shen, W. M. Shi, X. P. Yang, and B. X. Ye, "Particle swarm algorithm trained neural network for QSAR studies of inhibitors of platelet-derived growth factor receptor phosphorylation," *European Journal of Pharmaceutical Sciences*, vol. 28, no. 5, pp. 369–376, 2006.

[21] M. G. H. Omran, A. P. Engelbrecht, and A. Salman, "Dynamic clustering using particle swarm optimization with application in unsupervised image classification," *Proceedings of World Academy of Science Engineering and Thechnology*, vol. 9, pp. 199–204, 2005.

[22] Y. Kao, E. Zahara, and I. Kao, "A hybridized approach to data clustering," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1754–1762, 2008.

[23] T. Niknam and B. Amiri, "An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis," *Applied Soft Computing*, vol. 10, no. 1, pp. 183–197, 2010.

[24] C. D. Bocaniala and J. Sa da Costa, "Tuning the parameters of a fuzzy classifier for fault diagnosis. Particle swarm optimization versus genetic algorithms," in *Proceedings of the 1st International Conference on Informatics in Control, Automation and Robotics (ICINCO '04)*, pp. 157–162, Setubal, Portugal, August 2004.

[25] H. X. Pan, J. Y. Huang, H. W. Mao, and Z. W. Liu, "Fault characteristic extracting based on PSO," in *Proceedings of the 12th International Conference on Intelligent Engineering Systems (INES '08)*, pp. 139–144, February 2008.

[26] Q. F. Ma, *Gearbox fault diagnosis research based on particle swarm optimization neural network [M.S. thesis]*, North university of China, Taiyuan, China, 2006.

[27] X. F. Wang, J. Qiu, and G. J. Liu, "Discrete particle swarm optimization algorithm for gearbox fault symptom selection," *Journal of Aerospace Power*, vol. 20, no. 6, pp. 969–972, 2005.

[28] J. Y. Tham, L. Shen, S. L. Lee, and H. H. Tan, "A general approach for analysis and application of discrete multiwavelet transforms," *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 457–464, 2000.

[29] J. S. Geronimo, D. P. Hardin, and P. R. Massopust, "Fractal functions and wavelet expansions based on several scaling functions," *Journal of Approximation Theory*, vol. 78, no. 3, pp. 373–401, 1994.

[30] X. G. Xia, J. S. Geronimo, D. P. Hardin, and B. W. Suter, "Design of prefilters for discrete multiwavelet transforms," *IEEE Transactions on Signal Processing*, vol. 44, no. 1, pp. 25–35, 1996.

[31] V. Strela, P. N. Heller, G. Strang, P. Topiwala, and C. Heil, "The application of multiwavelet filterbanks to image processing," *IEEE Transactions on Image Processing*, vol. 8, no. 4, pp. 548–563, 1999.

[32] H. Matuyama, "Diagnosis algorithm," *Journal of JSPEl*, vol. 75, no. 3, pp. 35–37, 1991.

[33] P. Chen, T. Toyota, and Y. Sasaki, "Fuzzy diagnosis and fuzzy navigation for plant inspection and diagnosis robot," in *Proceedings of the IEEE International Conference on Fuzzy Systems (IEEE/IFES '95)*, pp. 185–193, Yokohama, Japan, March 1995.

[34] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, Calif, USA, 1972.

[35] Y. Shi and R. C. Eberhart, "Empirical study of particle swarm optimization," in *Proceedings of the Congress on Evolutionary Computation*, pp. 1945–1949, Washington, DC, USA, 1999.

[36] Y. H. Shi and R. C. Eberhart, "A modified particle swarm optimizer," in *Proceedings of the IEEE International Conference on Evolutionary Computation (ICEC '98)*, pp. 69–73, Anchorage, Alaska, USA, May 1998.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 236574, 9 pages
http://dx.doi.org/10.1155/2014/236574

Research Article

Two-Dimensional Impact Reconstruction Method for Rail Defect Inspection

Jie Zhao,[1,2] Jianhui Lin,[1] Jinbao Yao,[3] and Jianming Ding[1]

[1] *State Key Laboratory of Traction Power, Southwest Jiaotong University, Chengdu 610031, China*
[2] *Department of Industry Manufacture, Chengdu University, Chengdu 610106, China*
[3] *State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing 400030, China*

Correspondence should be addressed to Jie Zhao; zhaojie0111@163.com

Received 2 May 2014; Accepted 27 June 2014; Published 17 July 2014

Academic Editor: Xuefeng Chen

The safety of train operating is seriously menaced by the rail defects, so it is of great significance to inspect rail defects dynamically while the train is operating. This paper presents a two-dimensional impact reconstruction method to realize the on-line inspection of rail defects. The proposed method utilizes preprocessing technology to convert time domain vertical vibration signals acquired by wireless sensor network to space signals. The modern time-frequency analysis method is improved to reconstruct the obtained multisensor information. Then, the image fusion processing technology based on spectrum threshold processing and node color labeling is proposed to reduce the noise, and blank the periodic impact signal caused by rail joints and locomotive running gear. This method can convert the aperiodic impact signals caused by rail defects to partial periodic impact signals, and locate the rail defects. An application indicates that the two-dimensional impact reconstruction method could display the impact caused by rail defects obviously, and is an effective on-line rail defects inspection method.

1. Introduction

After the railroad was put into formal operation, the rail track inspection and maintenance became the key factor directly relating to the train operating safety. As the rail is used for long, some rail defects such as irregularity, fatigue crack, and fissure and pitting corrosion will appear [1, 2]. With increased traffic at higher speed, the rail defects will lead to fiercer vertical vibration impact of the vehicle-track coupling system [3–5], which may cause frequent failures of train and tracks, or even occurrence of fatal accidents. Therefore, from the perspectives of train operating safety and the efficiency of inspection and maintenance, it is necessary to investigate the on-line inspection methods for rail defects.

The traditional rail inspection techniques mainly include the magnetic particle inspection method and electromagnetic inspection method, which use the specific railway inspection vehicle to detect rail defects. These methods are of low efficiency and high cost and are unable to achieve on-line inspection while the train is operating. The current inspection methods for rail detects are primarily the image inspection method [6–9], ultrasonic inspection method [10, 11], signal processing method [12–16], and so on. The image inspection method is to detect the defects by analyzing the acquired images of the rail surface. Being direct and reliable, but due to large number of data acquired and slow processing speed, this method is unsuitable for full coverage inspection on the railroad track. The ultrasonic inspection refers to the inspection of rail defects by transmitting ultrasound to the rail and using the receiver to receive the reflected or diffused ultrasonic energy. This technique can detect fast but has high requirement for accuracy in equipment installation and the equipment would be easily interfered with high false alarm probability when the train moves. And the signal processing method finds out and locates the rail defects through making analysis of the vertical vibration signals of the vehicle-track

coupling system. It has the advantages of the abovementioned two techniques and can achieve the on-line inspection, so it becomes a focus for research.

Traditional signal processing methods are used to analyze the steady or approximate steady-state signals, for example, time domain statistics analysis, spectrum analysis, correlation analysis, and higher order spectral analysis; however, they are not applicable to analysis of the non-Gaussian and nonstationary signals caused by vertical vibration in vehicle-track coupling system. Modern signal processing methods such as short-time Fourier transform, wavelet transform, and empirical mode decomposition are well capable of analyzing the abovementioned signals but have unavoidable problems like difficulty in distinguishing the impacts caused by noise, aperiodic impact caused by train, and aperiodic impact caused by track.

After studying the wireless sensor network for the on-line inspection of rail defects, a two-dimensional impact reconstruction method is proposed in which multisensor information is taken as research target, analyzed and processed by improved modern time-frequency analysis method, and the image fusion processing technology is used to blank the periodic impacts caused by the defects of locomotive running gear or rail joints in the time-frequency spectrum of vertical impact signals, highlight and extract the impacts caused by rail defects, and realize the on-line inspection of rail defects.

2. Wireless Sensor Network for On-Line Inspection of Rail Defects

According to the actual distribution of train monitoring points, the wireless sensor network technology widely used in environmental monitoring [17] and military scouting [18] is introduced to the rail defect inspection. A wireless sensor network is constructed for on-line inspection of rail defects and realizes the real-time monitoring, sensing, and acquisition of vertical vibration signals of trains.

The sensor node distribution for rail defect inspection is designed as shown in Figure 1. The sensor at each node chooses the vibration impact sensor vertically installed above the axle box of the train bogie for monitoring the vertical vibration impact mainly resulting from track, wheels and transmission pair, and so forth.

All monitoring points distributed in Figure 1 may fully cover all monitored carriages. The vibration impact caused by the transmission failure and the flaws in wheels can transmit from axle, bearing, and axle box to the sensor node, and that caused by rail defects may also transmit to the sensor node through wheel, axle, and bearing and axle box. Besides, this distribution plan may enable each sensor node to gather the strong coupling data distinguishing wheel defect, transmission pair failure, and rail defect.

Since there are differences and certain coupling in the multisensor information collected by the wireless sensor network for the on-line inspection of rail defects, in order to make better fusion of the multisensor information, the monitoring points of all carriages at the same position will be deemed as a signal group unit to classify all sensor nodes in Figure 1, as shown in Table 1.

Table 1: Signal group classification.

Group number	Left side node number/right side node number
G_{L1}/G_{R1}	$L_1, L_5, \ldots, L_{N-3}/R_1, R_5, \ldots, R_{N-3}$
G_{L2}/G_{R2}	$L_2, L_6, \ldots, L_{N-2}/R_2, R_6, \ldots, R_{N-2}$
G_{L3}/G_{R3}	$L_3, L_7, \ldots, L_{N-1}/R_3, R_7, \ldots, R_{N-1}$
G_{L4}/G_{R4}	$L_4, L_8, \ldots, L_N/R_4, R_8, \ldots, R_N$

In this table, to distinguish the sensor nodes and signal groups at both sides of train, L_i denotes the ith sensor node at the left side, R_i represents the ith sensor node at the right side, G_{Lk} refers to the kth signal group at the left side, and G_{Rk} indicates the kth signal group at the right side.

The on-line inspection of rail defects requires all sensor nodes to collect data by equidistant space sampling, that is, sample once when the wheels turn for $1/z$ round. Let the driving distance of train be s, the signal collected by L_i sensor node can be represented by $l_i(s)$, and the signal collected by R_i sensor node is denoted by $r_i(s)$.

3. Rail Defect Inspection Based on the Two-Dimensional Impact Reconstruction

For the signals collected by a single-sensor node as mentioned above, the impacts caused by rail defects and the noises of locomotive running gear are not periodic, and those caused by rail joints and defects of locomotive running gear are periodic, so it is difficult to distinguish periodic and aperiodic impacts of tracks and vehicles. To avoid limitations of single-sensor information studying, a two-dimensional impact reconstruction method is put forward from the perspective of multisensor information fusion [19, 20]. This method is built on the wireless sensor network for on-line inspection of rail defects with the signal groups for study (see Table 1 for division of signal groups) and based on the multisensor information fusion of two-dimensional images. Its processing includes multisensor information preprocessing, shifting time-frequency analysis, and time-frequency spectrum image analysis and image fusion.

The wheel diameter is D, the full length of train (without external force, the distance of the internal side of coupler knuckle when the couplers at both ends are in closed position) is S_1, and the distance between rail joint gaps is S_2. For the convenience of explaining the two-dimensional impact reconstruction process, the signal group G_{Lk} is taken as an example for detailed discussion (the two-dimensional impact reconstruction of the signal group G_{Rk} is similar and thus omitted). The process is as follows.

3.1. Multisensor Information Preprocessing. Multisensor information preprocessing mainly comprises equidistant space resampling and threshold processing. First of all, all sensor node signals in the signal group for nonequidistant space sampling (hereinafter referred to as group signals) need the wheel speed collected by the speed sensor to make

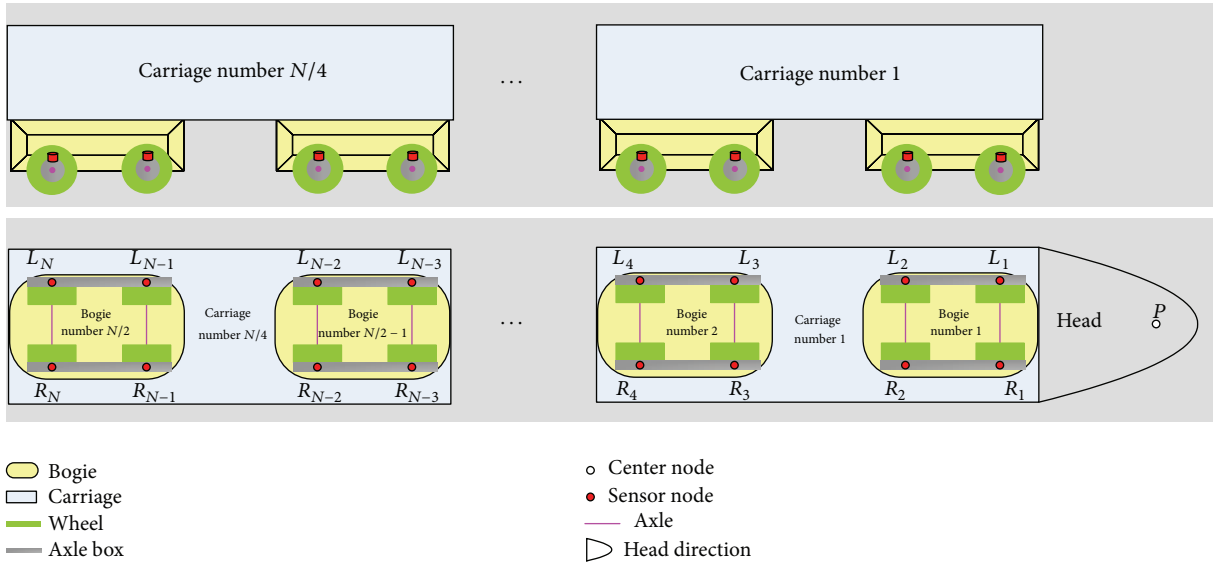


Figure 1: Node distribution shown in vertical and front views of the train.

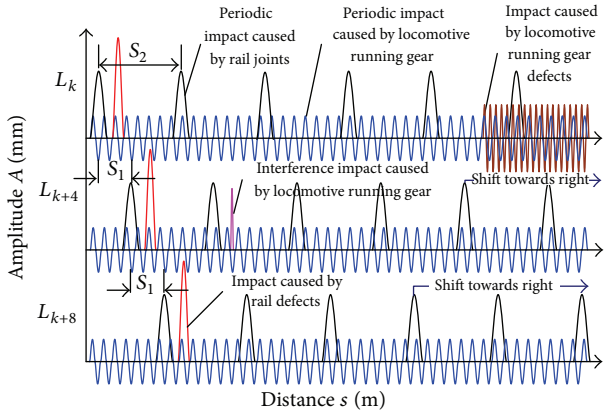


Figure 2: Sketch map of preprocessed inner group signal in space domain.

equidistant space resampling of signals and convert time domain signals to space domain signals and corresponding physical concepts such as time domain, period, and frequency to space domain, spatial period, and spatial frequency. Then, signals of sensor nodes for equidistant space sampling have been conducted the threshold processing, that is, artificially setting a threshold to filter the noise interference information the observer does not feel interested in.

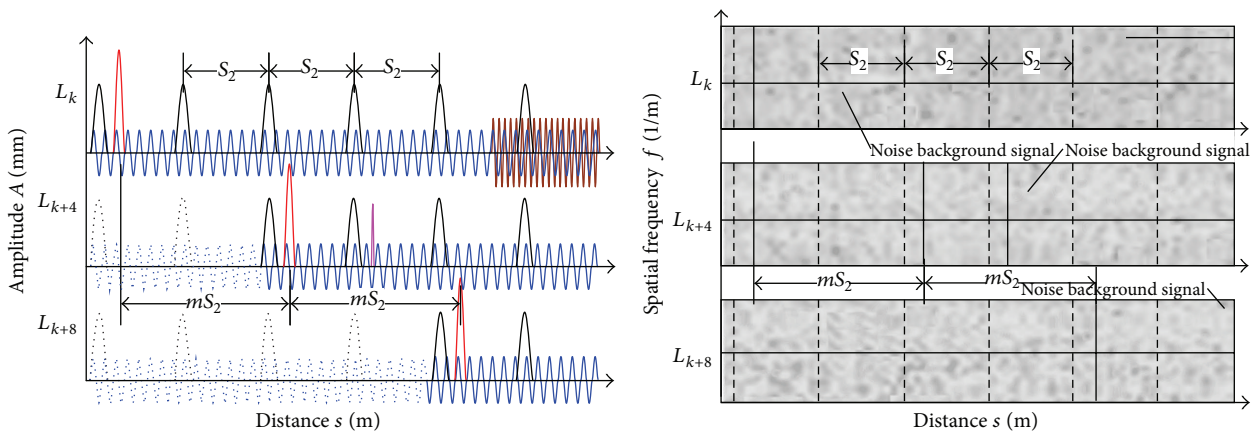
Figure 2 shows the waveforms of space domains obtained after preprocessing of signals gathered by the nodes L_k, L_{k+4}, and L_{k+8} in the signal group G_{Lk}, denoted by $l_k(s)$, $l_{k+4}(s)$, and $l_{k+8}(s)$, respectively. In order to simplify problem and better emphasize the key point, we suppose that $l_k(s)$, $l_{k+4}(s)$, and $l_{k+8}(s)$ all contain the spatial periodic impact (pulse type) caused by rail joints, spatial nonperiodic impact (pulse type) caused by rail defects, and spatial periodic impact (harmonic type) caused by locomotive running gear and noise signal (not drawn in the figure). Besides, let $l_k(s)$ contain the spatial periodic impact (pulse type) caused by the defect of locomotive running gear and let $l_{k+4}(s)$ contain the spatial interference impact (pulse type) caused by locomotive running gear.

3.2. Shifting Time-Frequency Analysis. Considering that the train in operation is influenced by various uncertain factors, the vertical vibration signals are obviously non-Gaussian and nonstationary. In the light of pertinence of signals in the group, for extracting the rail failure feature information, the shifting time-frequency analysis method is proposed. It is an improvement for the modern time-frequency analysis aiming at extracting of rail failures. This method aligns and shifts the preprocessed signals in groups in space domain to fully dig out the relevancy of signals in the groups and then acquires their spatial time-frequency spectrum by the modern time-frequency analysis, as shown in Figure 3.

Figure 3(a) describes the aligning and shifting of space domain signals shown in Figure 2. Firstly, it determines a benchmark, that is, the spatial periodic impact caused by rail joints in the signal $l_k(s)$ collected by the first node L_k in the signal group G_{Lk} as the benchmark. Secondly, the phase shifts towards left to return to zero, to be exact, the signal $l_i(s)$ collected by the node L_i is left shifted by $S_1(i-k)/4$ so that all node information in the signal group G_{Lk} is synchronized in distance. Finally, the phase shifts towards right to realize the spatial periodicity property; namely, the signal $l_i(s)$ is right shifted by $mS_2(i-k)/4$ so that the phase difference of spatial periodic impact caused by rail joints of signals $l_i(s)$ collected by the nodes is mS_2 and that of spatial nonperiodic impact caused by rail defects is also mS_2.

Figure 3(b) shows the spatial time-frequency spectrum of the signal shown in Figure 3(a).

In the following, we take short-time Fourier transform as an example of the modern time-frequency analysis, to deduce the formula of shifting time-frequency analysis.



(a) Aligning and shifting of space domain signals

(b) Time-frequency analyzing and processing

FIGURE 3: Sketch map of shifting time-frequency analyzing and processing.

After $l_i(s)$ is processed by using shifting time-frequency analysis, $L_i(s, f)$ can be obtained, as shown in

$$L_i(s,f) = \int_{-\infty}^{+\infty} l_i\left(s + \frac{(S_1 - mS_2)(i-k)}{4}\right) \times h(s-\tau)\, e^{-j2\pi f\tau} d\tau, \quad k = 1,2,3,4; \quad i = k, k+4, k+8, \ldots, N+k-4; \quad m = 2,3,\ldots, \tag{1}$$

where $L_i(s, f)$ is the spatial time-frequency spectrum after shifting time-frequency analysis and $h(s)$ represents the window function.

In the equation, the value of m must be the integral number greater than or equal to 2, or else the impact feature information may be blurred. Consider the following.

(a) If m is equal to 0, nonperiodic impacts caused by rail defects of nodes in the group have the same phase, which will blank the nonperiodic impacts after subsequent image fusion and result in loss of impact.

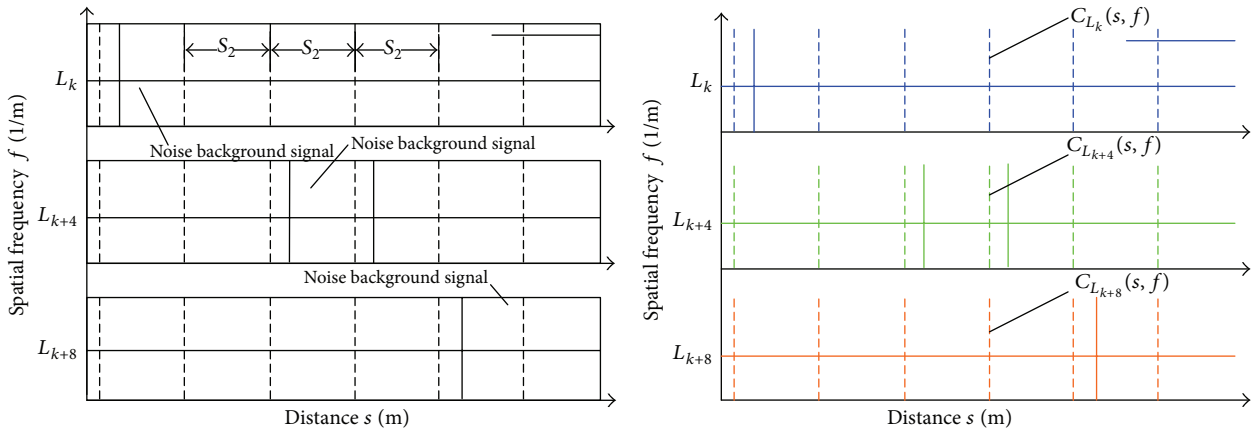
(b) If m is equal to 1, although nonperiodic impacts caused by rail defects of nodes in the group have the same phase difference between each other, after the subsequent image fusion, the nonperiodic impacts caused by rail defects have the same periodic features with the periodic impacts caused by rail joints, thus leading to impact overlap.

(c) If m is not an integral number, the periodic impacts caused by rail joints and locomotive running gear of nodes in the groups will have phase difference, and after the subsequent image fusion, the interference information may be introduced and the false impact may exist.

3.3. Time-Frequency Spectrum Image Analysis. In order to achieve better fusion performance of images in the subsequent "image fusion processing," the spatial time-frequency spectrum obtained after shifting time-frequency analysis still needs impact features extraction and node color labeling for further time-frequency spectrum image analysis, as shown in Figure 4.

3.3.1. Impact Feature Extraction. Impact feature extraction refers to a process in which the image grey scale processing technique is used to remove the noise background signals in order to highlight the impact signals. As the impact signals to be extracted from the spatial time-frequency spectrum contain the noise background signals, it is really necessary to perform the impact feature extraction. Its principle can be described as follows: firstly, the spatial time-frequency spectrum image is given statistics using the grey scale histogram, to find out the threshold between the impact signals and noise background signals, the dark grey corresponding to the impact signals and the light grey corresponding to the noise background signals. Then, using the captured threshold for gray threshold transform, the spatial time-frequency spectrum image is converted into the black and white binary images, the white corresponding to the noise background signals and the black corresponding to the impact signals. In the following, the impact feature extraction is in detail introduced with the schematic diagram.

The numeric area $[a, b]$ of the spatial time-frequency spectrum $L_i(s, f)$ as shown in Figure 3(b) turns into $[0, 255]$ when mapped onto the grey space. If the grey scale histogram is used for deciding the threshold λ_i, then the spectral value of the spatial time-frequency spectrum to which the threshold corresponds is $\sigma_i = (b-a)\lambda_i/255 + a$. After gray threshold transforming, all points which lie in $L_i(s, f) \geq \sigma_i$ on the spatial time-frequency spectrum are made black, and other points are made white, and the spatial time-frequency spectrum image is then converted into the black and white binary image, as shown in Figure 4(a). It can be seen that the noise background signal contained in the nodes of the signal group G_{Lk} is removed, and all impact signals are highlighted; thus the impact feature extraction is realized.



(a) Impact feature extracting

(b) Node color labeling

FIGURE 4: Sketch map of time-frequency spectrum image processing.

3.3.2. Node Color Labeling. Node color labeling refers to a process in which the blanking technique is used to remove the periodic impact signals from the impact signals arising from all kinds of factors, so as to extract the impact signals resulting from the rail defects. As the impact signals extracted are quite complex in terms of make-up, and, in particular, the periodic impact signals resulting from the locomotive running gear and rail joints, which are widely distributed in the whole space domain, may cause great effect on judgment of the impacts caused by the rail defects, it is really necessary to perform the node color labeling. If the spatial time-frequency spectrum of the signal of the node L_i within the signal group G_{Lk} is $L_i(s, f)$, then the spatial time-frequency spectrum of the node L_i can be labeled as follows: with the color labeling method as shown in (2) and its constraint equation (3), the different colors $C_{Li}(s, f)$ are used to label the spatial time-frequency spectrum image of the signal of the node L_i, as shown in Figure 4(b), so that the impact signals of the nodes at the same spatial position on the spectrum image are converted into the white background after the subsequent image fusion processing so as to blank the periodic impact signals. Consider

$$C_{Li}(s, f) = \begin{cases} \text{RGB}(255, 255, 255) & L_i(s, f) < \sigma_i \\ \text{RGB}(r_i, g_i, b_i) & L_i(s, f) \geq \sigma_i, \end{cases} \quad (2)$$

$$i = k, k+4, k+8, \ldots, N+k-4,$$

$$\sum_{i=k}^{N+k-4} r_i = 255 \quad 0 \leq r_i \leq 255,$$

$$\sum_{i=k}^{N+k-4} g_i = 255 \quad 0 \leq g_i \leq 255,$$

$$\sum_{i=k}^{N+k-4} b_i = 255 \quad 0 \leq b_i \leq 255,$$

$$\text{RGB}(r_i, g_i, b_i) \neq \text{RGB}(r_j, g_j, b_j) \quad i \neq j,$$

$$k = 1, 2, 3, 4; \quad i, j = k, k+4, k+8, \ldots, N+k-4, \quad (3)$$

where $L_i(s, f)$ is the spatial time-frequency spectrum of the signals of the node L_i within the signal group $L_i(s, f)$, r_i, g_i, and b_i are the color values of the components of the tricolor RGB, respectively, $C_{Li}(s, f)$ is the image color of the labeled $L_i(s, f)$, and σ_i is the spectral value of the spatial time-frequency spectrum to which the threshold λ_i decided by using the grey scale histogram corresponds.

Equation (2) defines the method of node color labeling, and (3) is its constraint. The first three equations in (3) are the necessary conditions for blanking, to make sure periodic impact caused by rail joints at each node within the group and periodic impact caused by locomotive running gear are blanked to white in subsequent "image fusion processing." The last inequation in (3) is a constraint designed to prevent the subsequent "image fusion processing" from causing the node code information loss and resulting in multiple impacts lines with the same color.

3.4. Image Fusion Processing. The multisensor information preprocessing, shifting time-frequency analysis, and time-frequency spectrum image analysis mentioned above are the first-phase preparations for the two-dimensional impact reconstruction method, and they convert the signals from one-dimensional space domain signals (as shown in Figure 2) into two-dimensional image signals (as shown in Figure 4). To explore as much relevancy between the multisensor information as possible and realize the feature extraction of the rail defects, it is necessary to perform the image fusion processing for the two-dimensional image signals derived from the processing above. And the principle can be described as follows: according to the pixel point color fusion algorithm as shown in (4), the spectrum (as shown in Figure 5) is obtained by fusing all the spatial time-frequency spectrum images of the node signals within the signal group

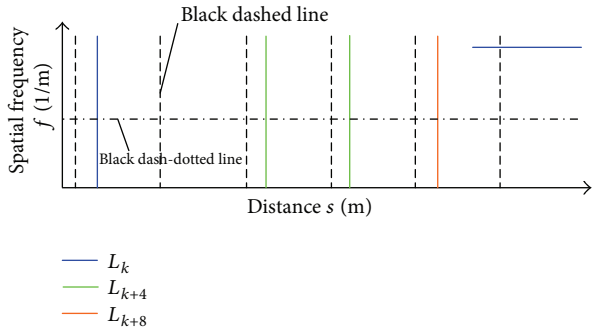


FIGURE 5: Sketch map of image fusion processing.

derived from the time-frequency spectrum image analysis, so as to realize the two-dimensional image reconstruction of the impact signals resulting from the rail defects. Consider

$$C_{G_{Lk}}(s,f) = \mathrm{RGB}(x_i, y_i, z_i),$$

$$x_i = 255 - \sum_{i=k}^{N+k-4} (255 - r_i) \,\%255,$$

$$y_i = 255 - \sum_{i=k}^{N+k-4} (255 - g_i) \,\%255, \quad (4)$$

$$z_i = 255 - \sum_{i=k}^{N+k-4} (255 - b_i) \,\%255,$$

$$k = 1, 2, 3, 4; \quad i = k, k+4, k+8, \ldots, N+k-4,$$

where r_i, g_i, and b_i are the color values of the components of the tricolor RGB, respectively, and $C_{G_{Lk}}(s,f)$ is the color of the two-dimensional impact reconstruction image fused by $C_{Li}(s,f)$ of all nodes within the signal group G_{Lk} on the left side of the train.

Equation (4) has the physical meaning as follows: when several images whose grey value is 255 are fused with one image whose grey value is h ($0 \le h \le 255$), the grey value of the image derived from such fusion will be h, and when several images whose grey value is 255 are fused, the grey value of the image derived from such fusion will be 255.

After substituting (2) and (3) into (4), the images of the nodes as defined in Figure 4(b) are fused, and the results of fusion are as shown in Figure 5. The figure shows that the periodic impacts caused by the rail joints and locomotive running gear (for the convenience of understanding and reading, in Figure 5, they are, respectively, expressed by the black dashed line and dash dot line, which are actually white dashed line and white solid line) are blanked after image fusion, which is because the above periodic impacts occur at the same place at each node within the signal group G_{Lk} and the color labeling at each node meets (3); then the color turns to RGB (255, 255, 255) after fusion. Meanwhile, the aperiodic impacts only occur at particular node within signal group G_{Lk} and the color at such node remains the same after fusion; as a consequence, the aperiodic impacts resulting from defects caused by the rail and the locomotive running gear are highlighted, and the node code information is retained by labeled colors.

After image fusion, the impact signals caused by the rail defects have been two-dimensionally reconstructed so as to make them different from other impact signals. The two-dimensional reconstructed impact signals caused by the rail defects have the following features.

(a) Impact signals cover all nodes, and $N/4$ impact lines appear.

(b) Impact lines vary in color, and all of them become white after image fusion.

(c) The period of the impact lines within partial space is mS_2.

Based on the feature information above, it can be judged whether or not there are rail defects, and if there are rail defects, they can also be located. When the three characteristics above occur, the rail can be judged defective and the defect is where the impact line at the first node L_k in signal group G_{Lk} is located.

4. Application Analysis

The two-dimensional impact reconstruction method focuses on multisensor information as study object while the modern time-frequency method takes single-sensor information as object, so the former is more suitable to inspect rail defects. To validate the effectiveness of the two-dimensional impact reconstruction method, short-time Fourier transform method and the two-dimensional impact reconstruction method have been used into the comparison and analysis of vibration data of a sixteen-carriage train.

Given that the train comprises 16 carriages, its whole length S_1 is 25,175 mm, and its wheel diameter D is 915 mm. The installation arrangement of the vibration impact sensors and speed sensors is as shown in Figure 1 (according to distribution of nodes in the signal group G_{L1}, 16 vibration impact sensors are installed above the axle boxes of the 16 carriages, separately). The data is captured by each vibration impact sensor once as the wheel rotates 1/2,000 cycle, and totally 16 groups of data are acquired (the train is running on the track whose length of rail S_2 is 25,000 mm while acquiring data).

Figure 6(a) shows the space domain waveform of vertical vibration signals at the node L_1. Figure 6(b) shows the result obtained after the vertical vibration signal is processed by short-time Fourier transform method (select Gaussian windowed short-time Fourier transform, and the size of Gaussian window is 12.8 mm). As shown in Figure 6(b), both the periodic impact (as shown by A in Figure 6(b)) and the aperiodic impact (as shown by B in Figure 6(b)) can be extracted effectively by short-time Fourier transform; however, it is difficult to tell whether B is caused by rail defects, for the reason that the aperiodic impact might be caused by noise, locomotive running gear, and so forth.

Figure 7 shows the detailed process in which the proposed two-dimensional impact reconstruction method is adopted to extract the rail defect features from the 16 groups

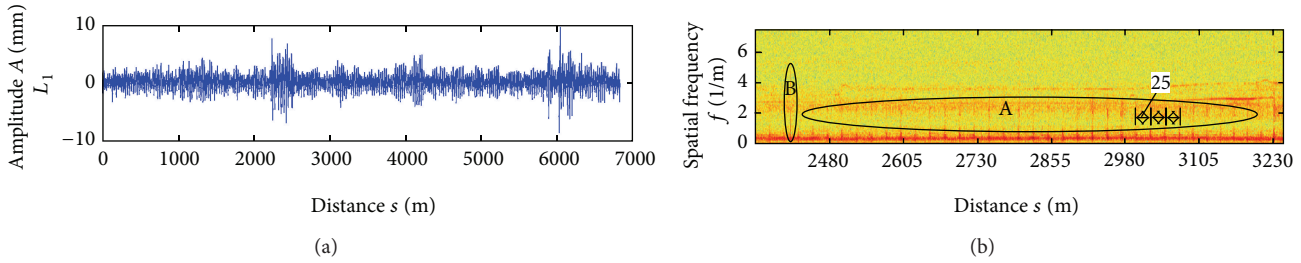


FIGURE 6: (a) Space domain wave of vertical vibration signals at the node L_1 and (b) spatial time-frequency spectrum after short-time Fourier transform.

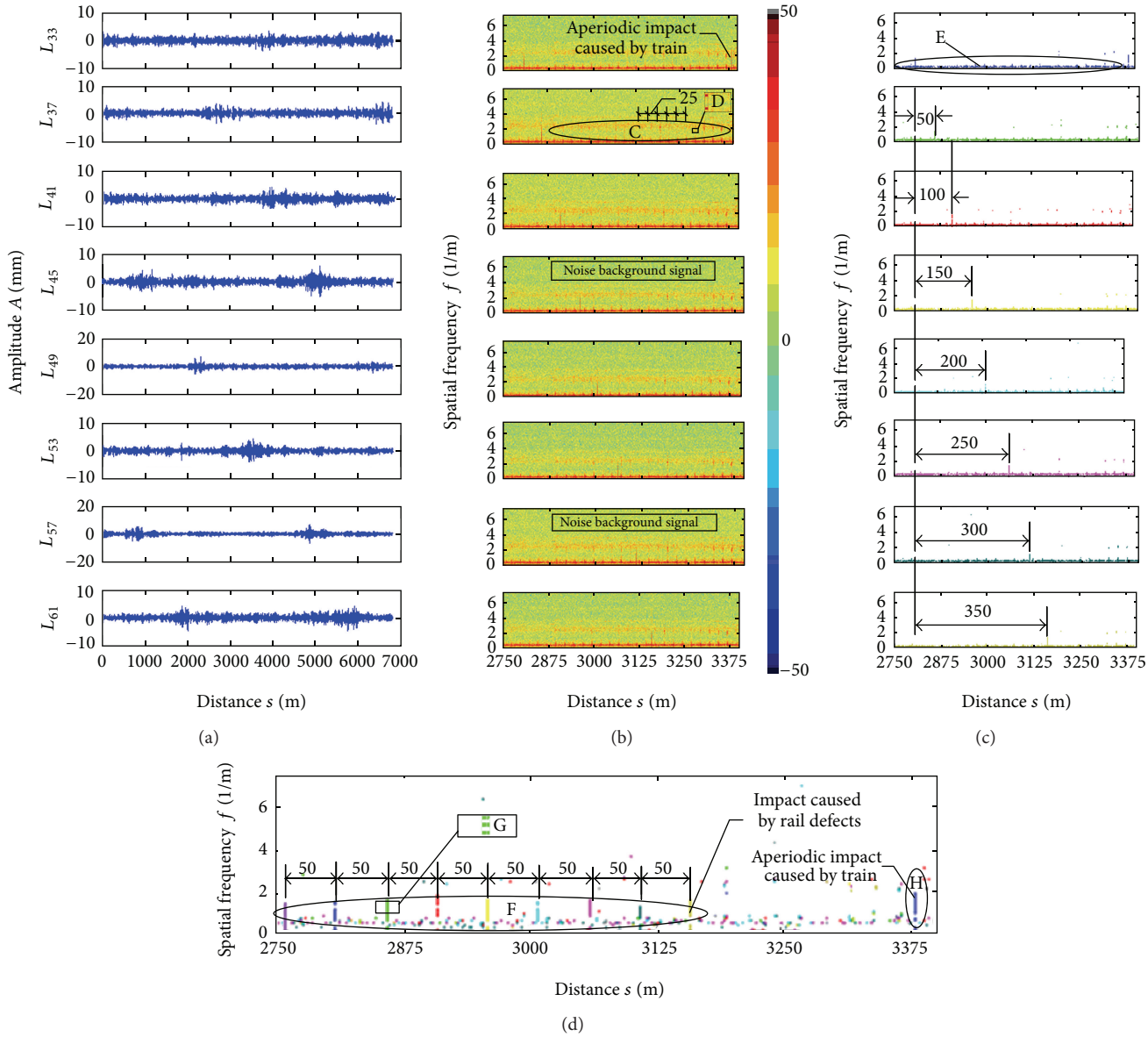


FIGURE 7: (a) Space domain wave of multisensor signal after preprocessing, (b) spatial time-frequency spectrum after shifting time-frequency analysis, (c) spatial time-frequency spectrum image after time-frequency spectrum image analysis, and (d) impact feature chart after image fusion processing.

of data corresponding to the signal group G_{L1} (only the last 8 groups of data are shown in the figure). Plots of the preprocessed 16 groups of original data in space domain are shown in Figure 7(a). Figure 7(b) shows the spatial time-frequency spectrum derived from the shifting time-frequency analysis (let $m = 2$, select Gaussian windowed short-time Fourier transform as the modern time-frequency analysis method, and the size of Gaussian window is 12.8 mm). Figure 7(c) shows the spatial time-frequency spectrum image derived from the time-frequency spectrum image analysis with the threshold $\sigma = 0.46$. Figure 7(d) shows the impact feature chart derived from the image fusion.

It can be seen from Figures 7(c) and 7(d) that the noise background signals in Figures 7(a) and 7(b) are largely removed, and the impact signals are highlighted. In Figures 7(b) and 7(c), the spatial frequency components of impact sequence whose period is $0.04\,\mathrm{m}^{-1}$ (corresponding to the periodic impacts arising from the rail joints, as shown by C and D in Figure 7(b)) are always seen at 25 m intervals across the whole travel, and the signals having the fixed spatial frequency of $0.35\,\mathrm{m}^{-1}$ across the whole travel (corresponding to the periodic impacts arising from the rotation frequency of the train wheel, as shown by E in Figure 7(c)) are all blanked to some extent, as shown in Figure 7(d). The spatial frequency components of impact sequence whose period is $0.02\,\mathrm{m}^{-1}$ (corresponding to the aperiodic impacts arising from the rail failure around 2,410 m, as shown by F and G in Figure 7(d)) are obviously seen at 50 m intervals across the partial travel, and a spatial wide band signal (corresponding to the aperiodic impacts arising from the train, as shown by H in Figure 7(d)) is seen around 3,395 m.

The application above shows that the two-dimensional impact reconstruction method is able to remove the noise background signals, highlight the aperiodic impact signals arising from the factors such as rail defects (as shown in Figure 7(d)), and allow the aperiodic impact signals resulting from rail defects to undergo the two-dimensional reconstruction to be different from the periodic impact signals within the partial space domain, so as to extract the impact features caused by the rail defects and realize the on-line inspection and location of the rail defects. In conclusion, two-dimensional impact reconstruction method is an effective method to extract track defects.

5. Conclusion

The two-dimensional impact reconstruction method integrates the wireless sensor network technique, inspection technique, information fusion technique, signal processing technique, and image processing technique and mainly involves such processes as multisensor information preprocessing, shifting time-frequency analysis, time-frequency spectrum image analysis, and image fusion processing. It can be used to make on-line analysis of the vertical vibration signals of the train as it is running, extract the aperiodic impact features caused by the rail defects effectively, and realize the on-line inspection and location of rail defects.

The two-dimensional impact reconstruction method mainly possesses the following characteristics.

(a) It uses the image grey scale processing technique, so it can remove the noise background signals and highlight all kinds of impact signals well.

(b) It uses the blanking technique, so it can remove the periodic impact signals from the impact signals well, and make the impact signals arising from the rail defects easily extracted.

(c) It can convert the aperiodic impacts arising from the rail defects into periodic ones within a certain partial space.

(d) It can on-line extract the impact features arising from the rail defects as the train is running and locate the rail defects.

The method also deserves further investigation on the following aspects.

(a) Considering that the proposed method involves shifting time-frequency analysis, image processing, and so forth, the processing speed is limited for the large amount of calculation. Although presently it can be used to give on-line inspection of the rail defects, it requires further study in terms of real-time inspection.

(b) Although it can remove the periodic impact signals, the impact signals still contain other aperiodic impact signals other than those arising from the rail defects (interference impacts caused by the locomotive running gear, impacts caused by defects of the locomotive running gear, etc.); further study is needed in this regard.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by the Key National Natural Science Foundation of China (Grant no. 61134002), and the National 863 plans projects (Grant no. 2011AA110501).

References

[1] E. G. Berggren, M. X. D. Li, and J. Spännar, "A new approach to the analysis and presentation of vertical track geometry quality and rail roughness," *Wear*, vol. 265, no. 9-10, pp. 1488–1496, 2008.

[2] M. J. M. M. Steenbergen, "Quantification of dynamic wheel-rail contact forces at short rail irregularities and application to measured rail welds," *Journal of Sound and Vibration*, vol. 312, no. 4-5, pp. 606–629, 2008.

[3] G. Lombaert and G. Degrande, "Ground-borne vibration due to static and dynamic axle loads of InterCity and high-speed trains," *Journal of Sound and Vibration*, vol. 319, no. 3–5, pp. 1036–1066, 2009.

[4] J. Vega, A. Fraile, E. Alarcon, and L. Hermanns, "Dynamic response of underpasses for high-speed train lines," *Journal of Sound and Vibration*, vol. 331, no. 23, pp. 5125–5140, 2012.

[5] X. Y. Liu and W. M. Zhai, "Analysis of vertical dynamic wheel/rail interaction caused by polygonal wheels on high-speed trains," *Wear*, vol. 314, no. 1-2, pp. 282–290, 2014.

[6] Q. Li and S. Ren, "A visual detection system for rail surface defects," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 42, no. 6, pp. 1531–1542, 2012.

[7] M. Guerrieri, G. Parla, and D. Ticali, "Mono and stereoscopic image analysis for detecting the transverse profile of worn-out rails," *Procedia—Social and Behavioral Sciences*, vol. 53, no. 1-2, pp. 611–621, 2014.

[8] L. S. Wu, Y. W. Li, H. W. Chen et al., "Research on rail defects automatic detection technology based on image region partition," *Laser Infrared*, vol. 42, no. 5, pp. 594–599, 2012.

[9] K. Xu, C. Yang, P. Zhou, and J. Liang, "3D detection technique of surface defects for steel rails based on linear lasers," *Journal of Mechanical Engineering*, vol. 46, no. 8, pp. 1–5, 2010.

[10] Y. Fan, S. Dixon, R. S. Edwards, and X. Jian, "Ultrasonic surface wave propagation and interaction with surface defects on rail track head," *NDT and E International*, vol. 40, no. 6, pp. 471–477, 2007.

[11] S. Mariani, T. Nguyen, R. R Phillips et al., "Noncontact ultrasonic guided wave inspection of rails," *Structural Health Monitoring*, vol. 12, no. 5-6, pp. 539–548, 2013.

[12] B. M. Hopkins and S. Taheri, "Track health monitoring using wavelets," in *Proceeding of the ASME Rail Transportation Division Fall Technical Conference (RTDF '10)*, pp. 9–15, October 2010.

[13] F. Lanza di Scalea and J. McNamara, "Measuring high-frequency wave propagation in railroad tracks by joint time-frequency analysis," *Journal of Sound and Vibration*, vol. 273, no. 3, pp. 637–651, 2004.

[14] A. Caprioli, A. Cigada, and D. Raveglia, "Rail inspection in track maintenance: a benchmark between the wavelet approach and the more conventional Fourier analysis," *Mechanical Systems and Signal Processing*, vol. 21, no. 2, pp. 631–652, 2007.

[15] L. Law, J. H. Kim, W. Y. H. Liew, and S. Lee, "An approach based on wavelet packet decomposition and HilbertHuang transform (WPDHHT) for spindle bearings condition monitoring," *Mechanical Systems and Signal Processing*, vol. 33, pp. 197–211, 2012.

[16] S. Jiang, Q. Fu, and Z. Wen, "Application of wavelet transform to obtain track static power spectrum density," *Journal of Traffic and Transportation Engineering*, vol. 4, no. 2, pp. 33–39, 2004.

[17] G. D. Zhou and T. H. Yi, "Recent developments on wireless sensor networks technology for bridge health monitoring," *Mathematical Problems in Engineering*, vol. 2013, Article ID 947867, 33 pages, 2013.

[18] K. Deng and Z. Liu, "Target tracking with dynamic clusters in wireless sensor network," *Acta Armamentarii*, vol. 29, no. 10, pp. 1197–1202, 2008.

[19] J. Zhang, B. Wang, J. Di, H. Yu, and B. Lu, "Research on information fusion for sensors multiple fault diagnosis," *Proceedings of the Chinese Society of Electrical Engineering*, vol. 27, no. 16, pp. 104–108, 2007.

[20] M. S. Safizadeh and S. K. Latifi, "Using multi-sensor data fusion for vibration fault diagnosis of rolling element bearings by accelerometer and load cell," *Information Fusion*, vol. 18, pp. 1–8, 2014.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 293878, 8 pages
http://dx.doi.org/10.1155/2014/293878

Research Article

A Fault Diagnosis Method for Rotating Machinery Based on PCA and Morlet Kernel SVM

Shaojiang Dong,[1,2] Dihua Sun,[1] Baoping Tang,[3] Zhenyuan Gao,[4] Wentao Yu,[5] and Ming Xia[6]

[1] *School of Mechatronics and Automotive Engineering, Chongqing Jiaotong University, Chongqing 400074, China*
[2] *School of Automation, Chongqing University, Chongqing 400044, China*
[3] *The State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing 400030, China*
[4] *Chongqing Academy of Metrology and Quality Inspection, Chongqing 401121, China*
[5] *School of Mechanical and Electronic Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China*
[6] *Mechanical and Electrical Engineering Department, Chongqing Vocational Institute of Safety & Technology, Wanzhou, Chongqing 404020, China*

Correspondence should be addressed to Shaojiang Dong; dongshaojiang100@163.com

Received 23 April 2014; Revised 8 June 2014; Accepted 15 June 2014; Published 8 July 2014

Academic Editor: Ruqiang Yan

A novel method to solve the rotating machinery fault diagnosis problem is proposed, which is based on principal components analysis (PCA) to extract the characteristic features and the Morlet kernel support vector machine (MSVM) to achieve the fault classification. Firstly, the gathered vibration signals were decomposed by the empirical mode decomposition (EMD) to obtain the corresponding intrinsic mode function (IMF). The EMD energy entropy that includes dominant fault information is defined as the characteristic features. However, the extracted features remained high-dimensional, and excessive redundant information still existed. So, the PCA is introduced to extract the characteristic features and reduce the dimension. The characteristic features are input into the MSVM to train and construct the running state identification model; the rotating machinery running state identification is realized. The running states of a bearing normal inner race and several inner races with different degree of fault were recognized; the results validate the effectiveness of the proposed algorithm.

1. Introduction

Rotating machinery is widely used in the modern factory. Unexpected mechanical faults could cause unscheduled downtime and loss. So, it is very important to diagnose the fault of the rotating machinery, to achieve effective fault diagnosis of the rotating machinery; firstly, the features should be extracted from the collected vibration data. Then, based on the extracted features an effective diagnosis model should be selected [1]. Feature extraction is the process of transforming the raw vibration data collected from running equipment to relevant information of health condition. There are three types of methods to deal with the raw vibration data: time domain analysis, frequency domain analysis, and time-frequency domain analysis. The three types of methods are often chosen to extract the feature. For example, Yan et al. [2] introduce that the time-frequency domain transform method wavelet is often used to describe the characteristics of the vibration signals. Gebraeel et al. [3] chose the average of the amplitudes of the defective frequency and its first six harmonics over time as the features. Yan et al. [4] chose the short-time Fourier transform to extract the features. Ocak et al. [5] chose the wavelet packet transform to extract the feature of bearing wear information. Because the time domain analysis and the frequency features from FFT analysis results often tend to average out transient vibrations and thus do not provide a wholesome measure of the bearing health status, in this paper, the time-frequency EMD is used to decompose the vibration signal and the EMD Shannon entropy is used to extract the original features from the signal.

Although the original features can be extracted, they are still with high-dimensional and include superfluous

information. So, the original features fusion and dimensional reduction method should be used to deal with the original features so as to select the typical features. The most commonly used features fusion and dimensional reduction method is principal component analysis (PCA). Sun et al. [6] used PCA to extract features from the run-to-failure test of vibration signals of bearings. Dong and Luo [7] proposed a PCA-based multivariate analysis method for bearing degradation process prediction. In this paper, the PCA is used to achieve the extraction of the most sensitive features.

After selecting the typical features, another challenge is how to achieve effective fault diagnosis of the rotating machinery based on the extracted features. The existing machinery fault diagnosis methods can be roughly classified into model-based (or physics-based modals) and data-driven methods. The model-based methods diagnosis the equipment fault using two models, the physical models based on the components and the damage propagation models based on damage mechanics. However, equipment dynamic response and damage propagation processes are typically very complex, and authentic physics-based models are very difficult to build [8]. Data-driven methods, known as artificial intelligent approaches, are derived directly from routine condition monitoring data of the monitored system, which achieve the fault diagnosis based on the learning or training process. The more prior the data used for the training process, the more accurate the model obtained. Artificial intelligent techniques have been increasingly applied to rotating machinery fault diagnosis recently. There have been some methods which are usually used for machinery fault diagnosis such as neural network and support vector machine (SVM) [9, 10]. However, the neural networks have the drawbacks of slow convergence; difficulty in escaping from local minima; and uncertain network structure, especially when doing the bearing fault diagnosis with large data. Those problems will be more troublesome. The SVM do not have those problems; however, the traditional SVM is not sensitive to the nonlinear feature classification, and, in recent years, the combination of wavelet theories and SVM has drawn considerable attention owing to its high classification ability for a wide range of applications and better performance than other traditional leaning machines. In this paper, the Morlet kernel is used to construct the new SVM model, and the PSO method is used to select the parameters [11].

The paper is organized as follows. In Section 2, the concept of EMD energy entropy is proposed and the EMD energy entropies of different vibration signals are calculated; the PCA is used to achieve the extraction of the most sensitive features. In Section 3, the Morlet wavelet kernel SVM model is presented. In Section 4, the running state identification model for rotating machinery fault diagnosis is applied to roller bearing. The conclusion of this paper is given in Section 5.

The flowchart of the proposed method is shown in Figure 1.

2. Methods of Signal Processing for Feature Extraction

This section presents a brief discussion on feature extraction from EMD. EMD is developed to decompose a signal into IMF components and every IMF has a unique local frequency. The IMF should satisfy two conditions. (1) In the whole data set, the number of extreme and the number of zero crossings must be either equal or different at most by one and (2) at any point, the mean value of the upper envelope and lower envelope is zero [12].

Once the extreme is identified, the maxima are connected by using the cubic spline and used as the upper envelope. The minima are interpolated as well to form the lower envelope. The upper and the lower envelopes should cover all the data in the time series. The mean of the upper and the lower envelope, $m_1(t)$, is subtracted from the original signal to obtain the first component $h_1(t)$ of the sifting process:

$$h_1(t) = x(t) - m_1(t). \tag{1}$$

Ideally, if $h_1(t)$ is an intrinsic mode function, the sifting process will stop. So, it will shift the signal again in the same way to get another component $h_2(t)$:

$$h_2(t) = h_1(t) - m_2(t), \tag{2}$$

where $m_2(t)$ is the mean of the upper and lower envelopes of $h_1(t)$.

Repeat steps until the residue satisfies some stopping criterion. The signal can be expressed as

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t), \tag{3}$$

where n is the number of IMFs, $r_n(t)$ is the residue which is a constant, a monotonic, or a function with only maxima and one minimum from which no more IMF can be derived, and $c_i(t)$ denotes IMF.

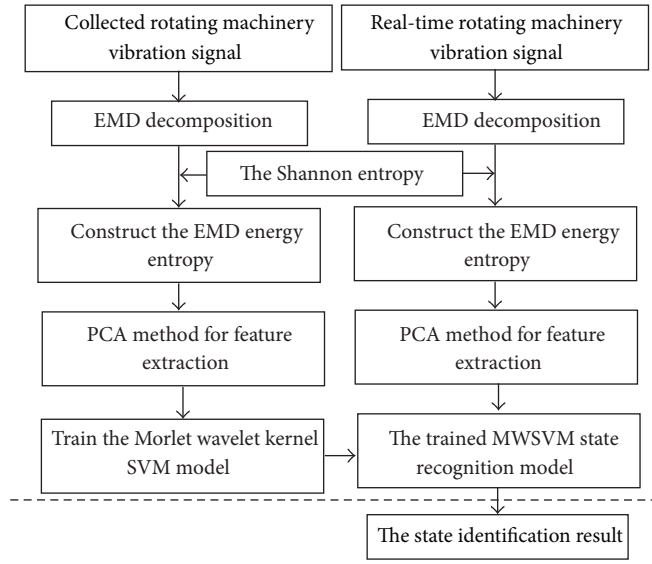
Once the n IMFs and a residue $r_n(t)$ are obtained, where the energy of the n IMFs $E_1; E_2; \ldots; E_n$ can be calculated, respectively, then, due to the orthogonality of the EMD decomposition, the sum of the energy of the n IMFs should be equal to the total energy of the original signal when the residue $r_n(t)$ is ignored. As the IMFs $c_1(t)$; $c_2(t); \ldots; c_n(t)$ include different frequency components, $\mathbf{E} = \{E_1, E_2, \ldots, E_n\}$ forms an energy distribution in the frequency domain of roller bearing vibration signal and then the corresponding EMD energy entropy is designated as

$$H_{\text{entropy}} = -\sum_{i=1}^{n} p_i \log p_i, \tag{4}$$

where $p_i = E_i/E$ is the percent of the energy of $c_i(t)$ in the whole signal energy ($E = \sum_{i=1}^{n} E_i$).

After the EMD energy entropy of the rotating machinery is calculated, the feature extraction method PCA is used to fuse the relevant useful features and extract the most sensitive features to work as the input of the proposed prediction model.

The procedure of feature extraction can be described as follows.



(a) The phase of model training (b) The phase of state recognition

Figure 1: The flowchart of the proposed method.

(1) Use the energy of the first five IMF components to get the features of the rotating machinery at each time.

(2) Use the PCA to reduce the original features dimensions and get one set of typical features as follows.

(a) Compute the covariance matrix from the data as

$$C = (X - \overline{x})(X - \overline{x})^T, \tag{5}$$

where X is the data matrix of EMD IMFs, N is the total number of patterns, and x represents mean vector of X.

(b) Compute the matrix of eigenvectors V and diagonal matrix of eigenvalues D as

$$V^{-1}CV = D. \tag{6}$$

(c) Sort the eigenvectors in V in descending order of eigenvalues in D and the data is projected on these eigenvector directions by taking the inner product in the data matrix sorted eigenvectors matrix as

$$\text{Projected data} = \left[V^T(X - \overline{x})^T\right]^T, \tag{7}$$

where V is of $n \times n$ dimension, and each row of it is an eigenvector. The features can be obtained.

(3) Use the features as input of the MSVM for rotating machinery fault state identification.

3. The Morlet Wavelet Kernel SVM Model

The support vector's kernel function can be described as the horizontal floating function, such as $k(x, x') = k(\langle x \cdot x' \rangle)$. In fact, if a function satisfies the condition of Mercer's theorem, it is the allowable support vector's kernel function. A specific Mercer's theorem description can be found in literature [13].

According to Mercer's theorem, the number of wavelet kernel functions which can be shown by the existent functions is few. Now, an existent wavelet kernel is given, the Morlet wavelet kernel. It can prove that this function can satisfy the condition of allowable support vector's kernel function. The Morlet wavelet function is defined as follows:

$$\psi(x) = \cos(w_o x)\, e^{-x^2/2}. \tag{8}$$

The Morlet wavelet kernel function is defined as follows:

$$\begin{aligned} k(x, x') = k(x - x') &= \prod_{i=1}^{d} \psi\left(\frac{x_i - x_i'}{a_i}\right) \\ &= \prod_{i=1}^{d} \cos\left[w_o\left(\frac{x_i - x_i'}{a_i}\right)\right] e^{\left[-(x_i - x_i')^2/2a_i^2\right]}. \end{aligned} \tag{9}$$

Then, the Morlet wavelet kernel function is being used as the support vector's kernel function, and the SVM is defined as

$$\begin{aligned} f(x) = \operatorname{sgn}\Bigg\{ &\sum_{i=1}^{n} a_i y_i \prod_{j=1}^{d} \cos\left[w_o\left(\frac{x_j - x_j'}{a_j}\right)\right] \\ &\times e^{\left[-(x_j - x_j')^2/2a_j^2\right]} + b \Bigg\}. \end{aligned} \tag{10}$$

Through (9) and (10), the Morlet wavelet kernel SVM is constructed, and the new constructed SVM which is effective in classification is used to achieve bearing running state recognition.

3.1. The Morlet Kernel SVM Parameters Selection. The particle swarm optimization algorithm (PSO) is used to select the SVM parameters, and the PSO was first proposed in 1995. It is an optimization method based on a set of particles whose coordinates are potential solutions in the search space. Particles in PSO will change their coordinates (their solutions) by migration. During migration, each particle adjusts its own coordinates based on its own past experience and other particles' past experiences.

The PSO was chosen to optimize the Morlet kernel SVM parameters through the following formula:

$$v_{ij}(t+1) = wv_{ij}(t) + c_1 r_{1j}\left(p_{ij}(t) - x_{ij}(t)\right) + c_2 r_{2j}\left(p_{g_j}(t) - x_{ij}(t)\right), \tag{11}$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1), \tag{12}$$

where the subscript "i" represents the ith particle. "j" represents the j-dimensional.

The subscript "t" represents the t generation. $v_{ij}(t)$ is the velocity of the ith particle in the tth iteration; $x_{ij}(t)$ is the position of the ith particle; $p_{ij}(t)$ is the pbest position of the ith particle; p_{g_j} is the gbest position (pbest represents the local optimum of the particles; gbest represents the overall situation optimum of the particles). The w represents the inertia weight. c_1, c_2 are learning factors. $r_1 \sim U(0,1)$, $r_2 \sim U(0,1)$ represent two independent random functions.

The process of optimizing the parameters γ, σ based on the PSO is given as follows.

(1) At the beginning of the optimization process, randomly initialize population sizes, c_1, c_2, ω, rand(1), and rand(2), determine the termination condition, positions, and velocities of the particle, mapping the Morlet kernel SVM parameters w_o, a into a group of particles, and initialize the initial position of each particle, pbest, gbest.

(2) When training the Morlet kernel SVM, use (11) as the PSO fitness function.

(3) Use the target parameters w_o, a as the particles, use their initial values as the LS-SVM parameters in step (2), and use the corresponding value of (11) as the optimal solution of the w_o, a.

(4) Use the initial error value of step (2) as the particle's initial fitness value and search the optimal value as the global fitness value among the initial fitness value and the corresponding particles as the current global optimal solution.

(5) Update the velocity and position vector.

(6) Resubstitute the updated parameters w_o, a into the Morlet kernel SVM model, retraining the Morlet kernel SVM model according to the step (2), save the output value, and calculate the fitness value of the particles again.

(7) Compare the saved global fitness value gotten in step (6) with the current particle's fitness value, and if the global fitness value is superior to the current particle's fitness value, update the current particle's fitness value according to step (5) and update the current particle's optimal value equal to the corresponding particle's optimal value gotten in step (6).

(8) While the termination conditions are not met, return to step (5).

(9) End the loop.

4. Validation

4.1. Case 1. In order to verify the effectiveness of the proposed method, bearing running state data sets of the normal state and several fault states were analyzed. The proposed method was applied to bearing fault signals obtained from the Case Western Reserve University [14]. The bearing type in the experiments is SKF 6205-2RS JEM. Experiments were conducted by using a 2 hp reliance electric motor. Bearings were seeded with faults by using electrodischarge machining. The test is to simulate the bearing normal running state and fault running states, with fault depths of 0.18 mm, 0.36 mm, 0.53 mm, and 0.71 mm at the inner raceway, outer raceway, and the ball to reflect the deteriorating state of the bearing; the inner-raceway fault signals were chosen in this case. Data was collected at the rate of 12,000 samples per second. 4096 data points were selected to analyze. 50 groups of test data of each fault state were selected, with 20 groups for training and the other 30 groups for testing. The collected vibration signals of normal state and inner-race four different fault depths are shown in Figure 2.

Next, the EMD decomposition was used to decompose each group of signals into IMFs, and Shannon entropy was used to extract the features. A group of inner-race entropy of Figure 2 is obtained, as shown in Table 1 (not normalized before).

Then, normalize the 20 groups of entropy values and input them into the PCA to reduce the dimension. In order to compare the dimension reduction and redundant treatment effect of PCA, the manifold learning method, local tangent space alignment (LTSA) [15], and the locality preserving projections (LPP) [16] method are used to reduce the dimension. The results are shown in Figures 3, 4, and 5. To be comparable, the dimensions of PCA, LTSA, and LPP are set to 3, so the input dimension of MWSVM is 3 and the neighborhood number is set to 10.

By comparing Figures 3, 4, and 5, the results show that the LTSA-based data dimension reduction method can not effectively separate the high-dimensional features, and there is still serious aliasing, which will affect the accuracy of the SVM state recognition effect. The LPP-based data dimension reduction method works better than the LTSA methods; however, there still have some features mixed together. The PCA-based data dimension reduction method can effectively separate the features of different running states with high calculation accuracy and a higher computational efficiency than the LPP and LTSA methods, which conform more to the actual project requirement. Thus, in the study, the PCA method is selected.

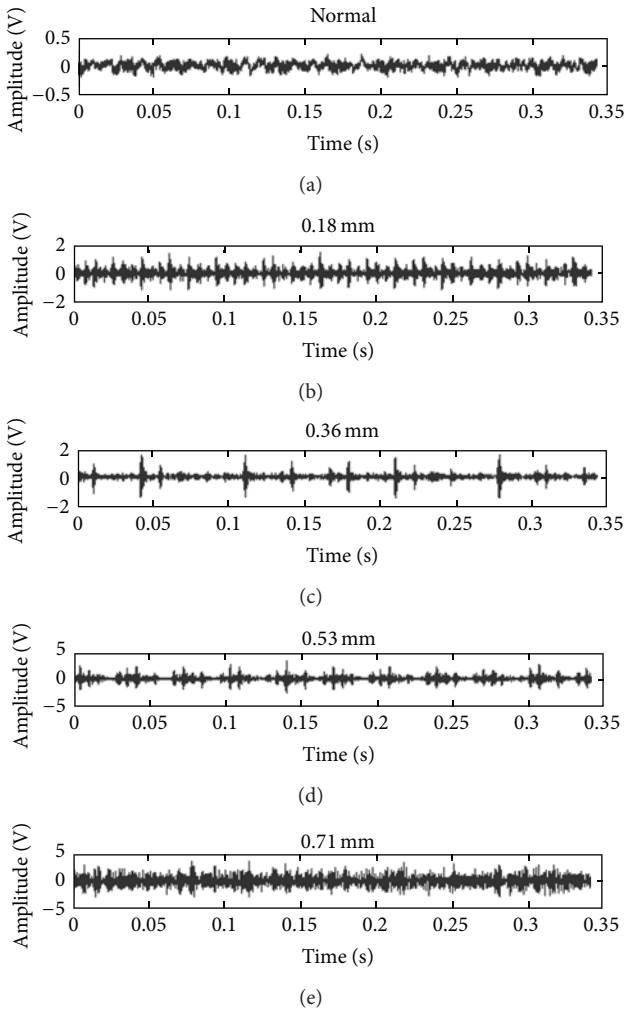


Figure 2: The collected vibration signals of normal state and inner-race four different fault depths.

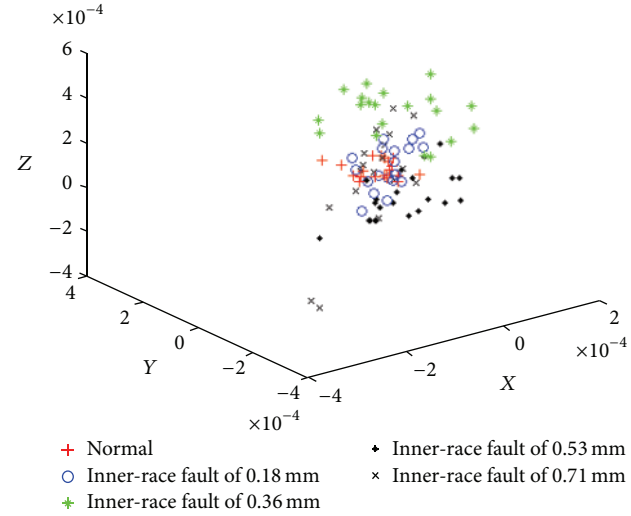


Figure 3: The dimension reduction and redundant treatment effect of LTSA.

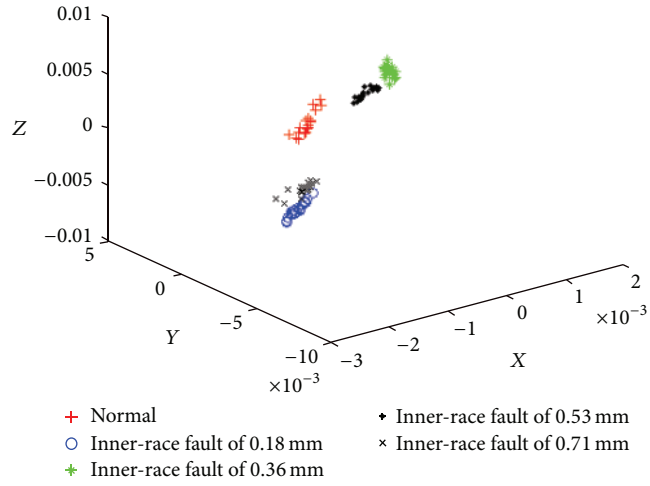


Figure 4: The dimension reduction and redundant treatment effect of LPP.

After dimension reduction with the PCA, the extracted features are input into the SVM to train the model so as to recognize the states. And the PSO is used to obtain the main parameters of the model, the particle swarm population size is set to 100, and the number of the particles is set to 20. The fitness function is set to get the minimum prediction error with the optimized parameters. The prediction error is set to 0.0001. The PSO particle's dimension is set to 2, the w is set to 0.5, the c_1 is set to 1, and the c_2 is set to 1. The optimized obtained parameter w_0 is 5, and a is 0.3. Then, the two parameters are used to build the Morlet kernel SVM model to train and predict the value. In order to compare the identifying effect with and without manifold learning method, the following comparisons are done.

(1) Use EMD Shannon entropy to extract the features and directly input the extract features into the MWSVM, without the PCA dimension reduction process.

(2) Use EMD Shannon entropy to extract the features and process the extracted features by LTSA to reduce the dimension and then input the features into the MWSVM.

(3) Use EMD Shannon entropy to extract the features and process the extracted features by LPP to reduce the dimension and then input the features into the MWSVM.

(4) Use the method proposed in this paper.

The comparison results are shown in Table 2.

Table 2 shows that, after the PCA-based dimension reduction method and features extraction, the accuracy of states recognition improved significantly, much higher than the other algorithms. Therefore, the use of PCA for dimension reduction in this research is necessary and valuable.

In order to further verify the identification accuracy of the proposed method, the features extracted by PCA are input into the neural network, traditional RBF SVM (with penalty factor C set to 100 and nuclear parameter γ set to 0.1), the

Table 1: A group of inner-race EMD energy entropy of different running states.

Running states	H_1	H_2	H_3	H_4	H_5
Normal state	1.3989	1.3798	1.3876	1.3946	1.3866
0.18 mm fault depth	1.1342	1.1153	1.1016	1.1276	1.1213
0.36 mm fault depth	0.8766	0.8547	0.8987	0.8895	0.8673
0.53 mm fault depth	0.6449	0.6451	0.6689	0.6783	0.6884
0.71 mm fault depth	0.49814	0.5121	0.5565	0.5223	0.5127

Table 2: The states recognition rate of three different methods (recognition rate η%).

States recognition methods	Normal state	0.18 mm fault depth	0.36 mm fault depth	0.53 mm fault depth	0.71 mm fault depth
Without PCA dimension reduction	91	61	76	83	81
Use the LTSA method dimension reduction	90	83	90	84	88
Use the LPP method dimension reduction	96	92	98	98	100
Use the PCA method dimension reduction	100	100	100	100	100

Table 3: The recognition rate of traditional RBF SVM and the MWSVM.

SVM type	Recognition rate η/%				
	Normal state	0.18 mm fault depth	0.36 mm fault depth	0.53 mm fault depth	0.71 mm fault depth
Neural network	90	96	94	93	90
RBF kernel	94	93	91	96	91
Symlet wavelet kernel	91	95	94	97	99
db wavelet kernel	90	98	96	92	90
Gaussian kernel	98	99	97	98	98
MWSVM	100	100	100	100	100

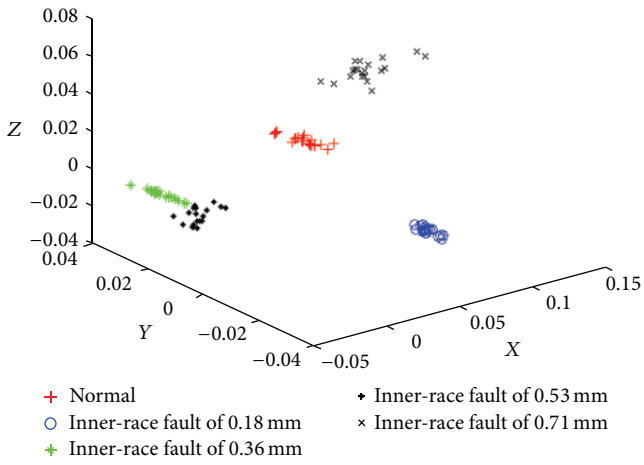


Figure 5: The dimension reduction and redundant treatment effect of PCA.

Symlet wavelet kernel SVM (with penalty factor C set to 100 and nuclear parameter γ set to 0.1), the db wavelet kernel SVM (with penalty factor C set to 100 and nuclear parameter γ set to 0.1), the Gaussian kernel SVM (with $a = 23.7$, penalty factor C set to 100, and nuclear parameter γ set to 0.1), and the MWSVM (with w_0 set to 5 and a set to 0.3). The comparison results are shown in Table 3.

Table 3 shows that the MWSVM can better identify and approach the sensitive features because of Morlet wavelet kernel. Thus, the choice of MWSVM to determine the bearing running states can effectively improve recognition accuracy.

Next, a comparison about the training and test time loss of different methods is implemented.

(1) The vibration data processed by EMD Shannon entropy and the extract features are directly input into the MWSVM, without the PCA dimension reduction.

(2) The vibration data processed by EMD Shannon entropy and the features are processed by PCA to reduce the dimension. Then, the extracted features are input into the RBF kernel SVM.

(3) The proposed method is in this research.

The comparison results are shown in Table 4.

In Table 4, after the dimension reduction, the recognition speed of SVM improved significantly. The time loss of the proposed method is the shortest. The reason is that the

Table 4: The time loss of three different methods.

Methods	Without PCA dimension reduction	By RBF kernel SVM	The proposed method
Time/s	133	79	58

Table 5: A group of EMD energy entropy of different running states of the actual signal.

Running states	H_1	H_2	H_3	H_4	H_5
Normal state	1.2657	1.2316	1.1091	1.1012	1.1316
Running for 2 months	1.0684	1. 0201	1.2452	1.2566	1.3910
Running for 4 months	0.9506	0.9896	0.9597	1.1166	0.9259
Running for 6 months	0.8923	0.8369	0.8949	0.8283	0.8675
Running for 8 months	0.6031	0.5108	0.6392	0.6967	0.7042
Running for 10 months	0.7663	0.7095	0.7953	0.7325	0.8034
Running for 12 months	0.7509	0.7993	0.8331	0.7408	0.8296

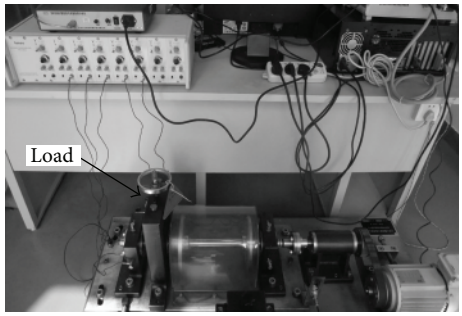


Figure 6: The test rig.

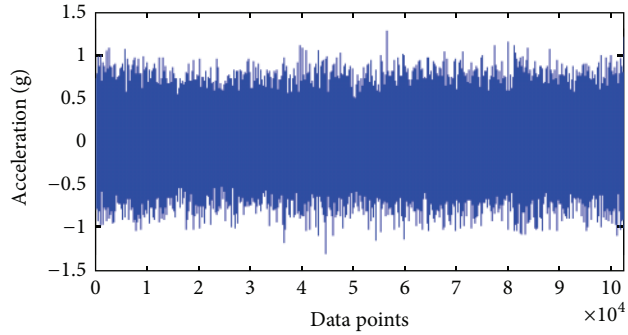


Figure 7: The collected vibration signal.

Morlet kernel is more sensitive to features classification and identification than the RBF kernel SVM. The result validates the proposed method and can effectively recognize the bearing running state.

4.2. Case 2. After validating the efficacy of the proposed method, the method is used on the actual application. The test rig is shown in Figure 6.

The bearings are hosted on the shaft; the shaft is driven by AC motor, the power is 0.55 KW, and the rotation speed is kept at 1000 rpm, with speed control and AC inverter controller. The brake maximum torque is 5 N·m, with a radial booster, using the magnetic clutch and brake. The rolling bearing is used, and a radial load of 29.4 N is added to the bearing. The data sampling rate is 25600 Hz and the data length is 102400 collected points, as shown in Figure 7. Every 2 hours, the vibration data is collected once. The bearing is run for one year. Then, a set of data from each of the 2 months is selected; the data sets are used to test whether or not the proposed method can identify the bearing running state. 4096 data points are selected to analyze, and 60 groups of collected data of different faults are obtained, with 30 groups for training and the other 30 groups for testing.

Next, each group of signals is decomposed by the EMD method, and the Shannon entropy is calculated. A group of features of different fault conditions are obtained, as shown in Table 5 (not normalized beforehand).

Table 6: The states recognition rate of different states based on the proposed method (recognition rate η%).

Running states	Recognition rate η%
Normal state	100
Running for 2 months	97
Running for 4 months	95
Running for 6 months	96
Running for 8 months	94
Running for 10 months	95
Running for 12 months	95

Then, the 30 groups' entropy values are normalized and input into the PCA in order to reduce the dimension and extract the typical features; the extracted features are input into the Morlet kernel SVM. The recognized results are shown in Table 6.

Table 6 shows that, although the actual bearing running state is very complex, the proposed method yields a high recognized accuracy. The results confirm that the proposed method can recognize the bearing running states effectively.

5. Conclusion

Firstly, this research used the EMD Shannon entropy method to extract the original features from the rotating machinery

vibration signals. The PCA was used to reduce the dimension and data redundancy of the entropy features. Through those methods, the typical features could be extracted effectively.

Then, in order to more accurately identify the bearing running state, the Morlet kernel was applied to construct the SVM recognition model; this can effectively improve the recognition accuracy of SVM.

Thirdly, through different comparisons, we can see that the proposed method makes good use of the advantage of all parts together to obtain better recognition accuracy and efficiency.

Finally, through the simulated signals and the tested signals in the research, the results show the significant efficacy of the proposed method in identifying the rotating machinery fault state.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research is supported by the Natural Science Foundation Project of CQ cstc2013jcyjA70896, Fundamental Research Funds for the Central Universities (Project no. CDJZR10118801). China Postdoctoral Science Foundation funded this research, Project no. 2014M552316. The authors are grateful to the anonymous reviewers for their helpful comments and constructive suggestions.

References

[1] S. Dong, B. Tang, and Y. Zhang, "A repeated single-channel mechanical signal blind separation method based on morphological filtering and singular value decomposition," *Measurement*, vol. 45, no. 8, pp. 2052–2063, 2012.

[2] R. Yan, R. X. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: a review with applications," *Signal Processing*, vol. 96, pp. 1–15, 2013.

[3] N. Gebraeel, M. Lawley, R. Liu, and V. Parmeshwaran, "Residual life predictions from vibration-based degradation signals: a neural network approach," *IEEE Transactions on Industrial Electronics*, vol. 51, no. 3, pp. 694–700, 2004.

[4] J. Yan, C. Guo, and X. Wang, "A dynamic multi-scale Markov model based methodology for remaining life prediction," *Mechanical Systems and Signal Processing*, vol. 25, no. 4, pp. 1364–1376, 2011.

[5] H. Ocak, K. A. Loparo, and F. M. Discenzo, "Online tracking of bearing wear using wavelet packet decomposition and probabilistic modeling: a method for bearing prognostics," *Journal of Sound and Vibration*, vol. 302, no. 4-5, pp. 951–961, 2007.

[6] C. Sun, Z. S. Zhang, and Z. J. He, "Research on bearing life prediction based on support vector machine and its application," *Journal of Physics*, vol. 305, Article ID 012028, 2011.

[7] S. Dong and T. Luo, "Bearing degradation process prediction based on the PCA and optimized LS-SVM model," *Measurement*, vol. 46, no. 9, pp. 3143–3152, 2013.

[8] Z. G. Tian, L. N. Wong, and N. M. Safaei, "A neural network approach for remaining useful life prediction utilizing both failure and suspension histories," *Mechanical Systems and Signal Processing*, vol. 24, no. 5, pp. 1542–1555, 2010.

[9] J. Lee, J. Ni, D. Djurdjanovic, H. Qiu, and H. Liao, "Intelligent prognostics tools and e-maintenance," *Computers in Industry*, vol. 57, no. 6, pp. 476–489, 2006.

[10] S. Dong, B. Tang, and R. Chen, "Bearing running state recognition based on non-extensive wavelet feature scale entropy and support vector machine," *Measurement*, vol. 46, no. 10, pp. 4189–4199, 2013.

[11] K. C. Gryllias and I. A. Antoniadis, "A Support Vector Machine approach based on physical model training for rolling element bearing fault detection in industrial environments," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 2, pp. 326–344, 2012.

[12] Y. Gan, S. Lifen, and W. Jiangfei, "An EMD threshold de-noising method for inertial sensors," *Measurement*, vol. 49, pp. 34–41, 2014.

[13] N. Cristianini and J. S. Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, New York, NY, USA, 2000.

[14] Case Western Reserve University Bearing Data Center, http://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website.

[15] Y. Zhan and J. Yin, "Robust local tangent space alignment via iterative weighted PCA," *Neurocomputing*, vol. 74, no. 11, pp. 1985–1993, 2011.

[16] F. Dornaika and A. Assoum, "Enhanced and parameterless Locality Preserving Projections for face recognition," *Neurocomputing*, vol. 99, pp. 448–457, 2013.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 537452, 10 pages
http://dx.doi.org/10.1155/2014/537452

Research Article

An Adaptive Maintenance Model Oriented to Process Environment of the Manufacturing Systems

Xun Gong,[1,2] Yixiong Feng,[1] Hao Zheng,[1] and Jianrong Tan[1]

[1] *State Key Lab of Fluid Power Transmission and Control, Zhejiang University, Hangzhou 310027, China*
[2] *Robotics and Microsystems Center, Soochow University, Suzhou 215006, China*

Correspondence should be addressed to Yixiong Feng; fyxtv@zju.edu.cn

Received 29 April 2014; Revised 28 May 2014; Accepted 30 May 2014; Published 29 June 2014

Academic Editor: Xuefeng Chen

We explored an adaptive maintenance model of the process environment to diagnose progressive faults in manufacturing systems. Progressive faults are usually caused by deterioration of the operating environment or aging and show stochastic properties. Many researchers have reported how to detect faults on the machine body in manufacturing systems. However, little research has been conducted on the process environment which causes progressive faults. To tackle this problem, we explored an adaptive maintenance model to detect progressive faults and repair the process environment on the E-repair location. When a difference of the environmental factor state is detected, it will combine the transcription factor and the state enzyme to locate fault source. Then the comprehensive maintenance program is derived to repair the operating environment while eliminating progressive faults. For the purpose of validation, this model was implemented on the process environment of the air separation plant. And the simulation experiments validated the feasibility and effectiveness of this method.

1. Introduction

Some large complex manufacturing systems are often operated under high pressures, at high temperatures, with fast material flows and complex manufacturing mechanism. In production, facility malfunction, environmental fluctuation, or feed stream instability can introduce a variety of process disturbances, which would aggravate the load of the equipment, accelerate wear and tear on the components, and increase the consumption of electricity or power. Severe combined disturbance propagations in a plant can be destructive. Obviously, such security threatening situations should be detected early, the potential impact on production should be precisely monitored, and operational solutions should be derived quickly.

Failure process in practical engineering applications mainly includes fault diagnosis and maintenance. Publishing of Beard's doctoral dissertation in 1971 marked the birth of fault diagnosis technology [1]. Since then, the fault diagnosis technology has become a research focus and scholars have conducted extensive and in-depth researchin two groups: (1) the model-based methods: some intelligent classification algorithms, such as artificial neural networks (ANNs) and support vector machines (SVM), have been successfully used for fault diagnosis of mechanical systems [2–5]. In machine condition monitoring and fault diagnosis, some researchers have used this as a tool for classification of faults [6, 7]; (2) the data-driven methods: this group of methods monitored and collected the input and output signals of the manufacturing process [8–11]. It extracted fault features from a large number of practical samples, described the relationships between faults and symptoms, and then constructed deep knowledge of expert systems [12, 13].

The concept of preventive maintenance (PM) involves the performance of maintenance activities prior to the failure of equipment [14, 15]. One of the main objectives of PM is to reduce the failure rate or failure frequency of the equipment. This strategy contributes to minimizing failure costs and machine downtime (production loss) and increasing product quality [16]. Reliability-centered maintenance emphasizes on equipment reliability and the consequences of equipment failure as the main basis for maintenance strategy [17, 18]; fault limited strategy decides whether to maintain the system or not by the failure rate and reliability as indicators [19];

condition-based maintenance strategies are monitoring the system [20, 21]; engineering systems maintenance strategy mainly considers the economic relations among the system devices [22].

The studies of environmental factors that affect product performance have focused on environmental simulation test before the operational process [23]. These methods exposed the defects of product components by the reliability enhancement testing. The adaptability of product was improved according to the scheduled test environment. The existing fault diagnosis and maintenance techniques are mostly oriented to device components or the system itself but not deep enough to the environmental factors stress which causes the failure [24].

From the year of failure statistics of the US airborne electronic equipment [25], it can be found that the fault caused by the temperature accounted for 22.2%; by the vibration accounted for 11.38%; by moisture accounted for 10%; by the dust accounted for 4.16%; by the salt spray accounted for 1.94%; by the impact accounted for 1.11%; by other causes accounted for 47.3%. From the above statistics it can be seen that the 52% of the total fault of the equipment system failures is caused by environmental stress factors of temperature, vibration, humidity, and pollution. At the same time the environmental stress factors also affect the validity of the detection data, thereby affecting the accuracy of fault diagnosis and blocking the maintenance work.

Motivated by those problems, we want to propose a maintenance method of the process environment to bridge the gap. The paper is presented as follows. Section 1 reviews briefly the development of the diagnosis and the maintenance. Section 2 presents the prerequisites of the method by analyzing the fundamental of the environmental stress response. Section 3 proposes the adaptive maintenance model to repair the abnormal process environment to be normal. Section 4 analyzes the temperature sensitivity of the air separation process (the precooling system/purification system/booster expansion turbine/refrigeration system). Section 5 presents the experiments on the air separation plant to test the feasibility and effectiveness of the adaptive maintenance model. Section 6 highlights findings of the paper and suggests potential research directions.

2. Prerequisite

Environmental stress response is defined as follows: during the P-F interval in the operating environment, it monitors the environmental factors of the equipment/system (the time from the potential failure to functional failure of the equipment called P-F interval). Once the early warning is in the potential failure state, the equipment/system is diagnosed timely to find out the disturbance source of the environmental factors. The operating environment is maintained in real time on the environmental repair location to avoid the duration of the potential failure state and the functional failure. It is shown in Figure 1.

By the above description of the fundamental of the environmental stress response and Figure 1, it can be seen that the presence, occurrence, diagnosis, and maintenance of

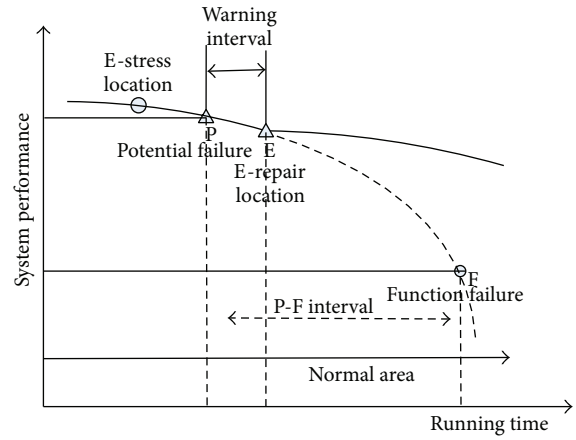


Figure 1: System performance of environmental stress response.

the operating environment potential failure have their own prerequisites. The adaptive maintenance model of the equipment/system operating environment in this paper requires some prerequisites like the following:

(1) a certain degree of fault sensitivity to environmental stress;

(2) determining an obvious potential failure state P;

(3) less than P-F interval time length of fault warning and carrying out the environmental stress response at the environmental restoration point E;

(4) the minimum P-F interval which must be long enough to arrange prevention and the environmental stress response in the potential failure process, but not the functional failure process.

The failure cumulative effect of the system is caused by environmental stress and it declines the system performance seriously. Various preventive techniques are used to diagnose potential failure timely to avoid the occurrence of functional failure. The environmental stress response on the environmental repair location repairs the operating failure environment to normal, to extend the lifetime of the system.

3. Methodology

3.1. Diagnostic Description. The mechanical device is constituted by the function, behavior, structure, carriers, and other design elements. In the mechanical system design theory, there are reciprocating mapping relationships among function domain, behavior domain, and carrier domain [26]. During operation, the device performance degrades because of fluctuations in E-factors. As shown in Figure 2, in order to diagnose the potential fault of the operating environment of the equipment in an abnormal operating environment, environment domain, monitoring domain, and state domain are increased beside equipment design elements domains.

Environmental domain is the set of E-factors in the equipment operating condition. The real-time state of the operating environment is gathered by monitoring the E-factors timely. The state data is traced back to the source of fluctuations in E-factors by diagnosing and analyzing.

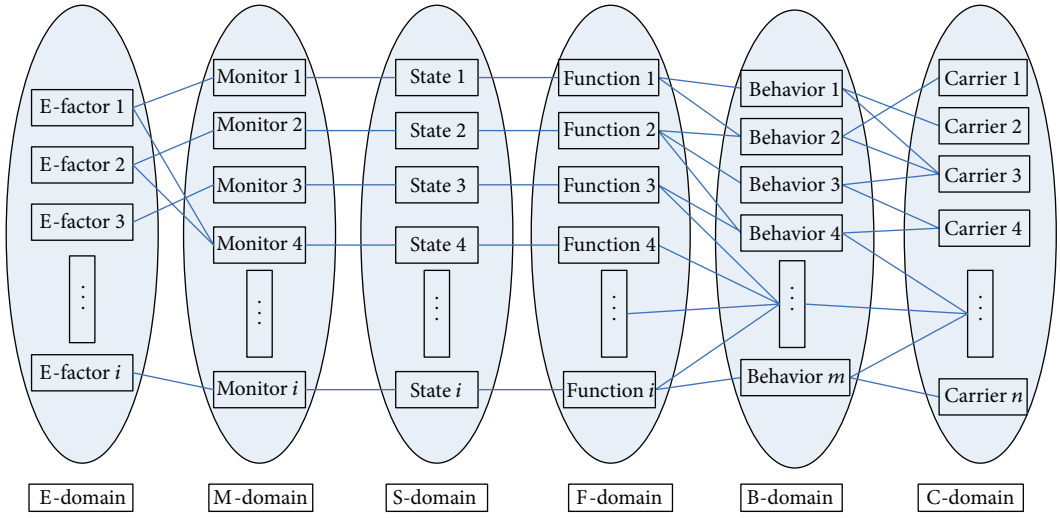


Figure 2: Domain structure mapping model.

The temperature is one of the E-factors which the equipment must face. Its fluctuations affect the performance of the system. For example, high temperature could cause thermal aging, structure changing, or physical expansion, while low temperature could cause material physical contraction, and temperature changes could cause the expansion and contraction, the institution stress, and so on.

Humidity is another E-factor. The high humidity stress can cause moisture accumulation and electrochemical reaction resulting in potential failures, while the low humidity stress can cause materials dried, grain, or other reactions.

The E-factors are various and random in reality, and there are mutual interactions among them. There are other E-factors leading to the system failure such as vibration/pressure/salt spray.

3.2. Adaptive Maintenance Model. The environment of the factory in which the equipment locates is extremely complex. There are a variety of disturbance sources causing the random fluctuations of the operating environment. The equipment is susceptible to progressive fault state under the E-factors stress during operation. If the abnormal operating environment is not diagnosed and maintained timely, the fault behavior is transferred and diffused among the mechanical components, and the progressive fault state also accumulates quickly. Eventually the function failure of the device would occur. Because of this, an adaptive maintenance model is proposed to diagnose and maintain progressive fault which is caused by environmental stress. It is shown in Figure 3. The adaptive maintenance model is composed of two parts. It is shown that the fluctuations of the manufacturing performance are caused by E-factors in part 1. And the monitoring and maintaining process are constructed in part 2.

The random fluctuations of the operating environment are considered during the system design process. E-factors are adjusted through the adaptive maintenance model based on the progressive fault state of the system. The method keeps the system in the normal operating environment.

Firstly, the system is running in the normal environment. According to the statistics, the stochastic dynamic environmental stresses are divided into several kinds such as temperature stress, humidity stress, vibration stress, and others.

Secondly, the stresses act on various components of the system causing a variety of physical and chemical reactions and affect the performance of the whole system to the progressive fault. If the progressive fault acts constantly on the system accumulating the fault state, that would lead to the function failure of the system. In the function failure process, the traditional fault diagnosis methods can be used for diagnosis and maintenance.

Thirdly, signal monitoring and recognition technology would be applied to the state of the system environment in the progressive fault process.

Fourthly, transcription factors and the array of state enzyme are constructed by the real-time operation of the system environment state and comapped to the expert system of the fault diagnosis established by the artificial intelligence methods. Environmental repair programs to the corresponding progressive fault states are obtained in the expert system.

Fifthly, the model uses transcription factors to activate or deactivate the components of the environmental repair programs and the state enzyme to adjust the trend and extent of the components. Then the closed-loop detection conditioning system is formed to repair the dynamic E-factors. And the progressive fault is removed by responding to the online environmental stress timely to maintain the normal operating environment of the system.

The specific steps are described as follows.

(1) Determine the system environment stress monitoring sites, L:

$$L = \{l_i \mid i = 1, 2, \ldots, n\}. \quad (1)$$

(2) The environment data of the system is obtained in the normal operation from the historical data statistics, T^0:

$$T^0 = \{t_i^0 \mid i = 1, 2, \ldots, n\}. \quad (2)$$

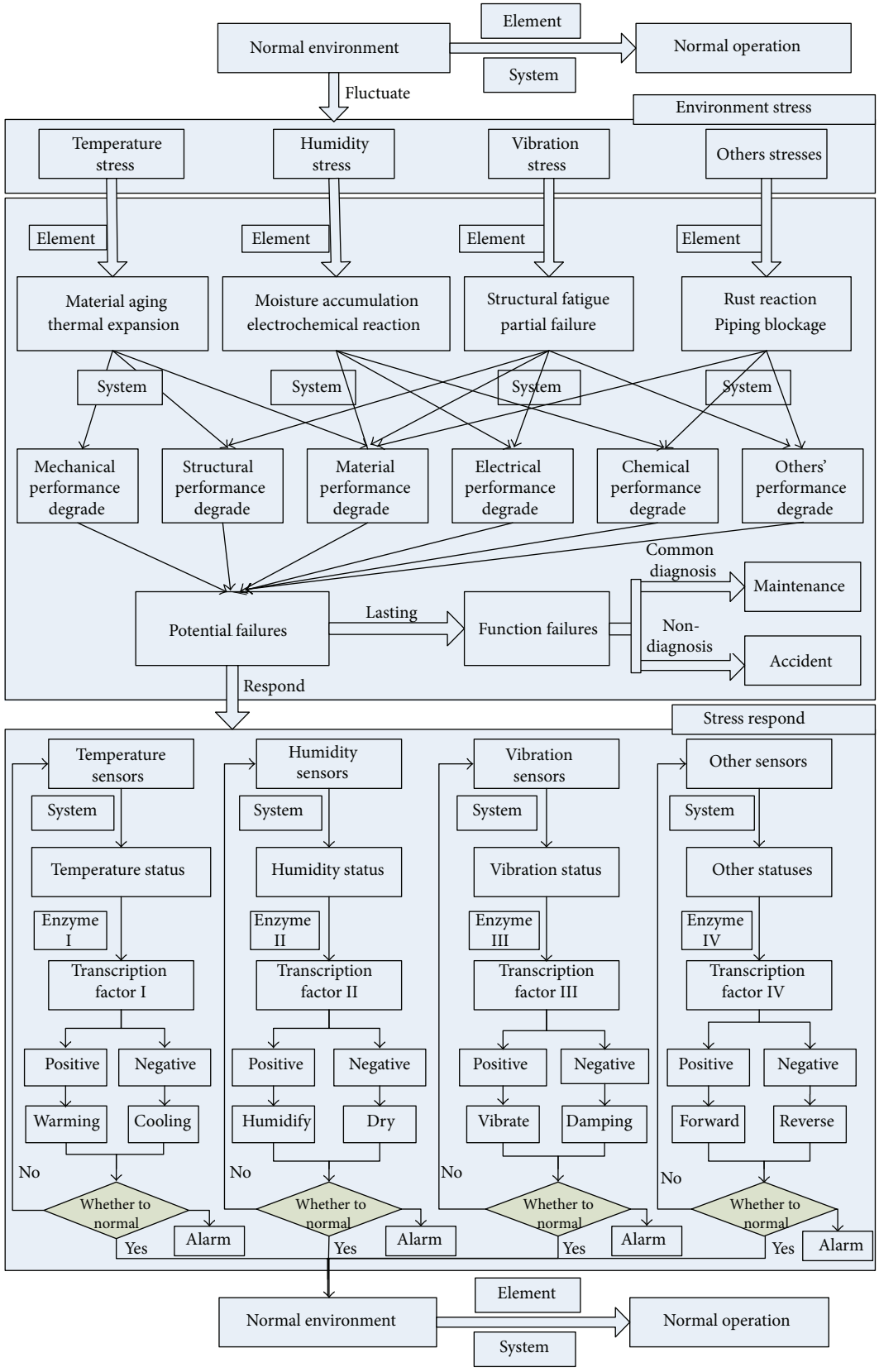


Figure 3: The adaptive maintenance model.

(3) The range of the environment factors of the system is obtained during the normal operation through the analysis of the system performance, T':

$$T' = \left\{ \left| t_i' \right| \mid i = 1, 2, \ldots, n \right\}. \quad (3)$$

(4) The state enzyme array is constructed based on the structure and the control parameters of the components of the environmental repair programs, E.

The state enzyme (e_{ij}) stands for the fluctuation extent of the failure/performance parameters of the same component of the system in different environments:

$$E = \left\{ e_{ij} = f\left(\Delta T_{ij}\right) \mid i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m \right\}. \quad (4)$$

(5) The system state array is constructed by real-time monitoring of the E-factors of the operation system, T:

$$T = \left\{ t_{ij} \mid i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m \right\}. \quad (5)$$

(6) The environmental stress transcription factor array is calculated by monitoring the E-factor state data, F:

$$F = \left\{ f_{ij} \mid i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m \right\},$$

$$f_{ij} = \begin{cases} 1 & \left|t_i'\right| < \left|t_{ij} - t_i^0\right|, t_{ij} < t_i^0 \\ 0 & \left|t_i'\right| > \left|t_{ij} - t_i^0\right| \\ -1 & \left|t_i'\right| < \left|t_{ij} - t_i^0\right|, t_{ij} > t_i^0 \end{cases} \quad (i = 1, 2, \ldots, n). \quad (6)$$

Transcription factor is one of the concepts of genetics. The transcription factor array can be used in the field of fault diagnosis to determine the locations of system components which could be affected by the environmental stress and in the progressive fault state. The positive or negative of the matrix values shows out the trend of the progressive fault.

(7) By the transcription factor array F and the state enzyme array E comapping to the expert system of the fault diagnosis, obtain the comprehensive maintenance program for the environmental stress response M. While the transcription factor value (f_{ij}) is 1, it shows that the monitoring site (l_i) is under the environmental stress and the response is positive, and while the transcription factor value (f_{ij}) is 0, it shows that the monitoring site (l_i) is not under the environmental stress and it needs no response, and while the transcription factor value (f_{ij}) is −1, it shows that the monitoring site (l_i) is under the environmental stress and the response is negative.

The repair effects of the environmental restoration program (M) is determined by comparing the environment state data of the operating system (such as T_i and T_{i+1}). If it is found out that the data and the trend do not match or exceed the regulatory range of the comprehensive maintenance (M), then the alarm would be worn. Once in this kind of situation, the components of the system should be diagnosed or the system should be upgraded to prevent functional failure happening.

Table 1: Monitoring data of the booster expansion turbine.

Contents	Status			
	1	2	3	4
Outlet temperature (°C)	−157.7	−161.7	−165.7	−167.7
Oxygen content in oxygen (%)	0.99953	0.99944	0.99933	0.99927
Nitrogen content in nitrogen (%)	0.9998634	0.9998636	0.9998638	0.9998639
The ratio of oxygen extraction (%)	0.6731	0.6840	0.6951	0.7007

4. Sensitivity Analysis

The air separation plant has the typical characteristics such as electrohydraulic system coupling, the complex spatial structure, and the high failure risk. The operation reliability and the product quality of the air separation plant are sensitive to environmental stresses such as temperature, humidity, vibration, and pressure. The temperature is one of the most important environmental stresses which the air separation plant must face. Its fluctuations affect the performance of the system.

The air separation process is mainly composed of a refrigerating system and rectification system, as shown in Figure 4. The simulation model includes the compressed air system, the precooling system, the purification system, the heat exchange system, the refrigeration system, and the distillation system. And there are parts of the air separation experimental setups shown in Figure 5 (the heat exchange system).

The precooling system of the air separation plant cools down by the circulating water. The temperature of the cooling water will decline with the drop of the atmospheric humidity and temperature and would enhance the cooling effect on the compressed air heat load; on the contrary, the temperature will rise with the rise of the atmospheric humidity and temperature and would decline the cooling effect.

The reduction of the cooling effect leads to the high temperature of the air before entering the purification system. And the high temperature air would affect the normal operation of the molecular sieve purification into the potential failure condition. Then the system production load is decreased, the productivity declines, and the power consumption increases.

85% to 90% of the cooling capacity is produced by the turbine expander of the full low pressure air separation plant. The purified air passes through the turbine expander cooling to form the raw solution. The raw solution is separated into gas products by the distillation column.

The air separation processes are simulated and the outlet temperatures of the booster expansion turbine are adjusted in the simulation. Collect the changes of the purity of the oxygen product and the nitrogen product, the oxygen extraction rate, and other data. Parts of the collected data are shown in Table 1 and Figure 6.

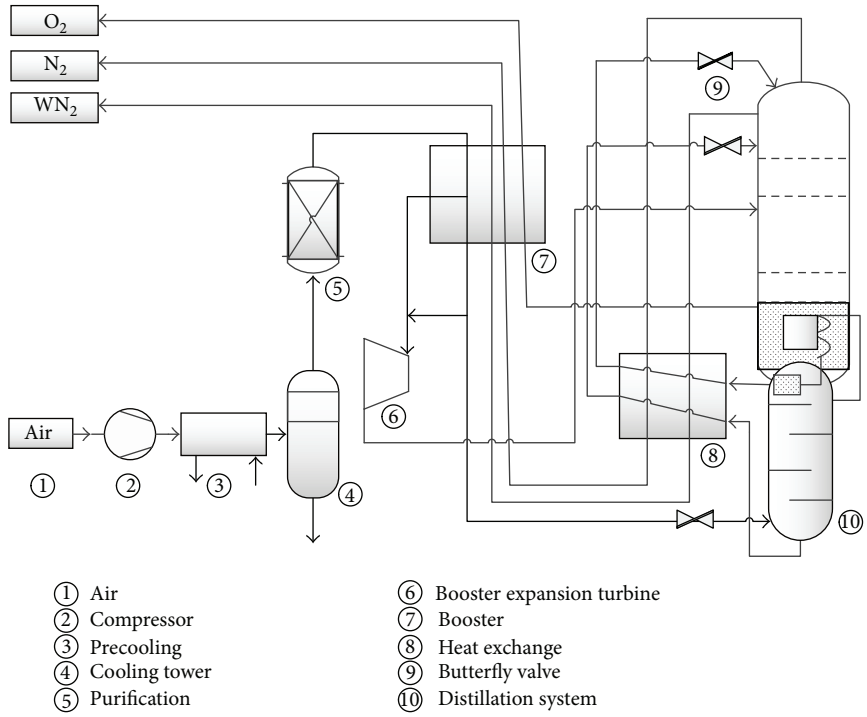


Figure 4: The simulation model of air separation process.

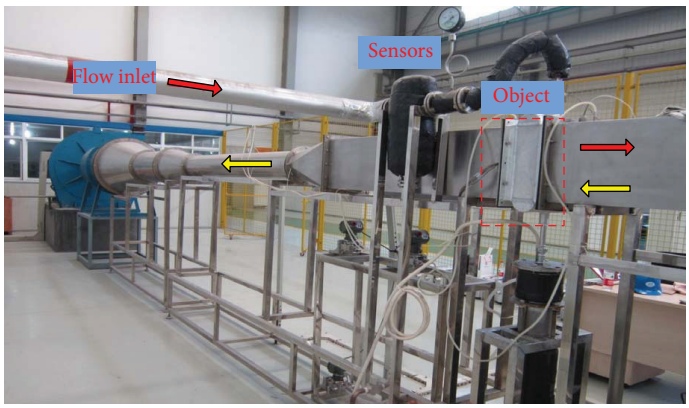


Figure 5: Parts of the air separation equipment.

It is shown directly in Figure 6. When the temperature of the booster turbo expander declines, the expansive air is blown into the feeding plate of the upper rectifying tower and its superheat drops down. That causes the vaporization of the reflux liquid in the rectifying section of the upper tower decrease. The liquid-gas ratio in the rectifying section is higher than at the original operational temperature. The oxygen fraction condenses fully from the vapor phases to the liquid phase because of the increase of the reflux liquid flow. The oxygen content of the vapor phase drops while the nitrogen content rises.

5. Experiments

A series of experiments were carried out to validate the proposed approach. The air separation process is simulated based on the data collected in the field on a certain type (7500/15000) of air separation plant in the normal operating environment. Construct the normal operating state (S_0) of the simulated air separation according to the requirements of the air separation such as the convergence and the thermal coupling of the tower systems. And then monitor the E-factors such as inlet-outlet temperatures, pressures, and flows of the key equipments such as air compressors, precooling systems, purification systems, and the booster expansion turbine.

With reference to the system historical monitoring data, the process is simulated by adjusting E-factors on the air separation system. The simulation results are shown in Table 2.

For the operations, select the temperature stress which is one of the E-factors as the object from Table 2. Detect mainly

Table 2: Statistics of the air separation process compared simulation.

Parameters	S	The operation in the stress state but without the environmental stress response				The operation in the stress state and with the environmental stress response			
		T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
Air compressor									
cp_in	T	41.3	41.3	41.3	41.3	41.3	41.3	41.3	41.3
cp_out	T	72.48	75.62	78.14	79.97	72.48	75.62	78.14	79.97
Precooling system									
c_in	T	72.57	75.47	77.98	80.51	72.57	75.47	77.98	80.51
c_out	T	18.88	19.56	20.17	20.85	18.88	17.32	15.41	13.82
Purification system									
p_in	T	18.96	18.81	19.19	19.91	18.96	17.29	15.27	13.68
p_out	T	29.87	31.56	34.24	37.11	29.87	25.75	21.29	17.14
Air booster									
pb_in	T	29.87	31.44	34.12	37.03	29.89	25.77	21.35	17.28
pb_out	T	96.2	98.4	102.3	105.7	96.2	82.7	75.3	65.8
Booster expansion turbine									
e_in	T	−106.68	−104.25	−102.67	−100.18	−106.68	−108.53	−110.24	−111.08
e_out	T	−158.2	−157.5	−155.7	−153.5	−158.2	−158.9	−159.8	−161.4
Distillation tower									
Flow	O	7395	7392	7388	7382	7395	7406	7419	7430
	N	14547	14542	14539	14537	14547	14561	14572	14581
	WM	11826	11829	11831	11837	11826	11821	11814	11809
Purity	O	0.999287	0.999311	0.999338	0.999354	0.999287	0.999012	0.998772	0.998616
	N	0.918654	0.918649	0.918642	0.918636	0.918654	0.918692	0.918706	0.918712

the inlet and outlet temperatures of the air compressor, the precooling system, the purification system, and the booster expansion turbine and monitor the components of the gas products from the rectification tower. Then construct the array (L) of the system monitoring locations:

$$L = \{\text{cp\_in, cp\_out, c\_in, c\_out, p\_in, p\_out, pb\_in, pb\_out, e\_in, e\_out}\}. \quad (7)$$

Construct the fluctuation range array (T') to maintain the healthy state based on the environment parameters of the operation system.

Consider $T' = \{40\ \ 60\ \ 60\ \ 4\ \ 4\ \ 6\ \ 6\ \ 30\ \ 10\ \ 3\}$. Construct the system state array (T) based on the simulated data in Table 2.

The monitoring data ($T_1 \sim T_4$) in Table 2 are obtained by the timing simulation of the air separation process when the system is in the stress state (S_8) but without the environmental stress response. It shows that the potential failure of the system accumulates constantly.

The monitoring data ($T_5 \sim T_8$) in Table 2 are obtained by the timing simulation of the air separation process when the system is in the stress state (S_8) and the environmental stress response.

Then $T_{10,4}$ and $T'_{10,4}$ can be obtained from Table 2:

$$T_{10,4} = \begin{vmatrix} 41.30 & 41.30 & 41.31 & 41.31 \\ 72.48 & 75.62 & 78.14 & 79.97 \\ 72.57 & 75.47 & 77.98 & 80.51 \\ 18.88 & 19.56 & 20.17 & 20.85 \\ 18.96 & 18.81 & 19.19 & 19.91 \\ 29.87 & 31.56 & 34.24 & 37.11 \\ 29.87 & 31.44 & 34.12 & 37.03 \\ 96.19 & 98.41 & 102.30 & 105.70 \\ -106.68 & -104.25 & -102.67 & -100.18 \\ -158.2 & -157.5 & -155.7 & -153.5 \end{vmatrix},$$

$$T'_{10,4} = \begin{vmatrix} 41.30 & 41.30 & 41.31 & 41.31 \\ 72.48 & 75.62 & 78.14 & 79.97 \\ 72.57 & 75.47 & 77.98 & 80.51 \\ 18.88 & 17.32 & 15.41 & 13.82 \\ 18.96 & 17.29 & 15.27 & 13.68 \\ 29.87 & 25.75 & 21.29 & 17.14 \\ 29.89 & 25.77 & 21.35 & 17.28 \\ 96.19 & 82.71 & 75.34 & 65.80 \\ -106.68 & -108.53 & -110.24 & -111.08 \\ -158.18 & -158.91 & -159.83 & -161.42 \end{vmatrix}. \quad (8)$$

Locations of system components which are in the potential failure can be detected in the positive stresses by

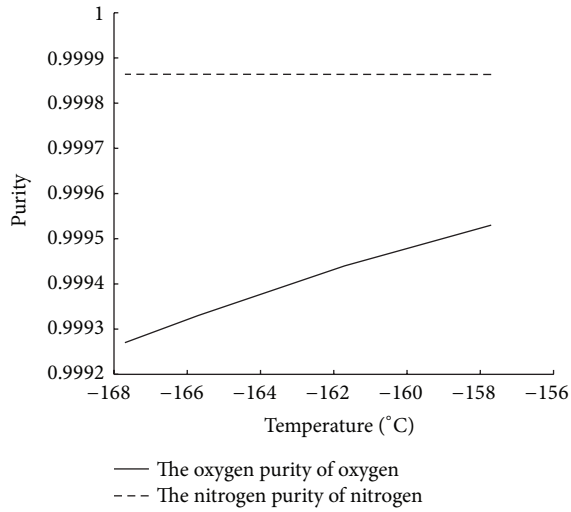


Figure 6: Diagrams of outlet temperatures of the booster expansion turbine and oxygen and nitrogen purity.

the transcription factor array (F_8). Start a negative regulation scheme for the temperature stresses to cool down or cooling the ambient system environment, while the state enzyme array (E_8) is obtained by comapping and calculating.

Make the most obvious state of the environmental stress (S_8) as the state of the potential failure which is under the temperature stress in computing of the environmental stress response. According to formula 8, the transcription factor array (F_8) can be obtained as follows:

$$F_8 = \begin{cases} 26.8 < 40 & & \longrightarrow & 0 \\ 36 < 60 & & \longrightarrow & 0 \\ 36 < 60 & & \longrightarrow & 0 \\ 5.98 > 4 & 12.2 < 18.18 & \longrightarrow & -1 \\ 6.06 > 4 & 12.2 < 18.26 & \longrightarrow & -1 \\ 14.07 > 6 & 15.8 < 29.87 & \longrightarrow & -1 \\ 14.07 > 6 & 15.8 < 29.87 & \longrightarrow & -1 \\ 37.12 > 30 & 59.08 < 96.2 & \longrightarrow & -1 \\ 6.93 < 10 & & \longrightarrow & 0 \\ 6.9 > 3 & -165.1 < -158.2 & \longrightarrow & -1 \end{cases} \quad (9)$$

$$F_8 = \{0, 0, 0, -1, -1, -1, -1, -1, 0, -1\},$$

$$E_8 = \left\{e_{ij} = f\left(\Delta T_{ij}\right) \mid i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m\right\}$$

$$= \{0, 0, 0, 1.49, 1.52, 2.35, 2.35, 1.24, 0, 2.3\}$$

Cool down the components which are under the potential failure and the ambient environment by operating the comprehensive maintenance program (M_8) to adjust the E-factors.

The comparisons of the air separation process simulation between those without (solid line) and those with (dotted line) the environmental stress response are shown in Figures 7 and 8. The outlet temperatures of the precooling system, the purification system, and the air compressor are shown in Figure 7. The inlet-outlet temperatures of the booster expansion turbine are shown in Figure 8.

It can be seen by the comparisons that it is effective to take the environmental stress response on the operational

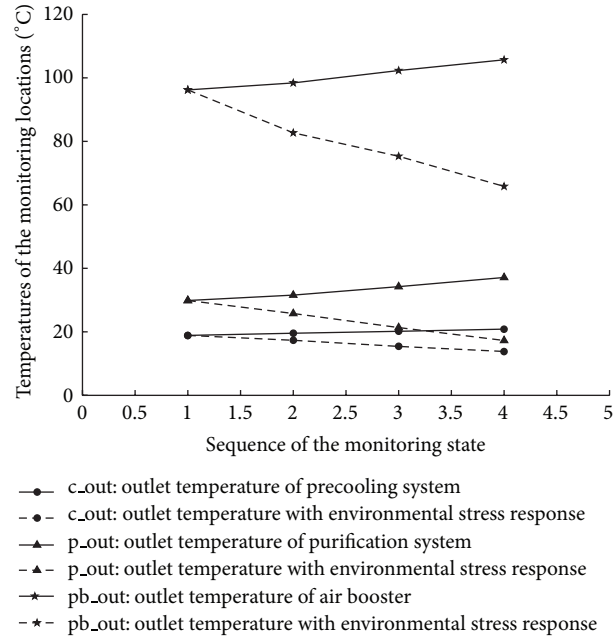


Figure 7: Contrast diagram of the running temperatures of the components.

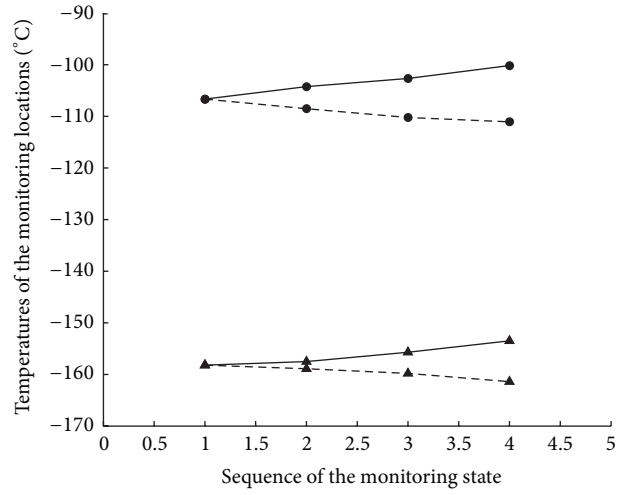


Figure 8: Contrast diagram of the inlet and outlet temperatures of the booster expansion turbine.

phase of the air separation system as shown in Table 3. The reducing of the temperature means low power consumption of the manufacturing system and the increasing of production means higher benefits. Also the normal operating environment would extend the service life of the system.

6. Conclusion

We presented an adaptive maintenance model to repair the process environment which caused progressive faults in

Table 3: Analysis of the effect of the adaptive maintenance model.

Locations	Status			
	T_4 (°C)	T_8 (°C)	ΔT (°C)	Δ%
c_out	20.85	13.82	−7.03	−33.72
p_out	37.11	17.14	−19.97	−53.81
pb_out	105.7	65.8	−39.9	−37.75
e_in	−100.18	−111.08	−10.9	−10.88
e_out	−153.5	−161.4	−7.9	−5.15
F_O	7382	7430	48	+0.65
F_N	14537	14581	44	+0.30

the air separation plant system. This maintenance approach includes the following. (1) The diagnostic model monitors the environmental states of the plants and also compares the inputs/outputs and presettings to detect faults. (2) The mapping structure is constructed with the I/O environmental states and behaviors of carriers, while the state enzyme and the transcription factor array are calculated through the expert system. (3) The comprehensive maintenance program is obtained by the comapping of the state enzyme and the transcription factor array for the environmental stress response.

For the future research, we suggest to optimize the deployment of the sensors for the model. Through preselection of sensor locations, it may improve the detection of the system with optimal cost and sensor configuration.

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 51322506 and 51175456), Zhejiang Provincial Natural Science Foundation of China (No. LR14E050003), the Fundamental Research Funds for the Central Universities, Innovation Foundation of the State Key Laboratory of Fluid Power Transmission and Control, and Zhejiang University K.P. Chao's High Technology Development Foundation. Sincere appreciation is extended to the reviewers of this paper for their helpful comments.

References

[1] R. V. Beard, *Failure Accommodation in Linear Systems through Self-Reorganization*, MIT, Cambridge, Mass, USA, 1971.

[2] M. Demetgul, "Fault diagnosis on production systems with support vector machine and decision trees algorithms," *The International Journal of Advanced Manufacturing Technology*, vol. 67, no. 9–12, pp. 2183–2194, 2013.

[3] J. Yang, Y. Zhang, and Y. Zhu, "Intelligent fault diagnosis of rolling element bearing based on SVMs and fractal dimension," *Mechanical Systems and Signal Processing*, vol. 21, no. 5, pp. 2012–2024, 2007.

[4] S. F. Yuan and F. L. Chu, "Support vector machines-based fault diagnosis for turbo-pump rotor," *Mechanical Systems and Signal Processing*, vol. 20, no. 4, pp. 939–952, 2006.

[5] A. Widodo and B. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 21, no. 6, pp. 2560–2574, 2007.

[6] A. Widodo, B. Yang, and T. Han, "Combination of independent component analysis and support vector machines for intelligent faults diagnosis of induction motors," *Expert Systems with Applications*, vol. 32, no. 2, pp. 299–312, 2007.

[7] Y. Yang, D. Yu, and J. Cheng, "A fault diagnosis approach for roller bearing based on IMF envelope spectrum and SVM," *Measurement: Journal of the International Measurement Confederation*, vol. 40, no. 9-10, pp. 943–950, 2007.

[8] P. Stepanic, I. V. Latinovic, and Z. Djurovic, "A new approach to detection of defects in rolling element bearings based on statistical pattern recognition," *International Journal of Advanced Manufacturing Technology*, vol. 45, no. 1-2, pp. 91–100, 2009.

[9] F. Pan, S. R. Qin, and L. Bo, "Development of diagnosis system for rolling bearings faults based on virtual instrument technology," *Journal of Physics: Conference Series*, vol. 48, article 467, 2006.

[10] C. Angeli, "An online expert system for fault diagnosis in hydraulic systems," *Expert Systems*, vol. 16, no. 2, pp. 115–120, 1999.

[11] C. Angeli and A. Chatzinikolaou, "On-line fault detection techniques for technical systems: a survey," *International Journal of Computer Science & Applications*, vol. 1, no. 1, pp. 12–30, 2004.

[12] B. Samanta and K. R. Al-Balushi, "Artificial neural network based fault diagnostics of rolling element bearings using time-domain features," *Mechanical Systems and Signal Processing*, vol. 17, no. 2, pp. 317–328, 2003.

[13] T. Lindh, *On the condition monitoring of induction machines [Ph.D. thesis]*, Lappeenranta University of Technology, 2003.

[14] I. B. Gertsbakh, *Models of Preventive Maintenance*, North-Holland Publishing, Oxford, UK, 1977.

[15] H. Löfsten, "Management of industrial maintenance—economic evaluation of maintenance policies," *International Journal of Operations and Production Management*, vol. 19, no. 7, pp. 716–737, 1999.

[16] J. S. Usher, A. H. Kamal, and W. H. Syed, "Cost optimal preventive maintenance and replacement scheduling," *IIE Transactions*, vol. 30, no. 12, pp. 1121–1128, 1998.

[17] L. Pintelon and G. Waeyenbergh, "A practical approach to maintenance modelling," in *Flexible Automation and Intelligent Manufacturing*, J. Ashayeri, W. G. Sullivan, and M. M. Ahmad, Eds., pp. 1109–1119, Begell House, New York, NY, USA, 1999.

[18] G. Waeyenbergh and L. Pintelon, "A framework for maintenance concept development," *International Journal of Production Economics*, vol. 77, no. 3, pp. 299–313, 2002.

[19] H. Wang, "A survey of maintenance policies of deteriorating systems," *European Journal of Operational Research*, vol. 139, no. 3, pp. 469–489, 2002.

[20] M. Wiseman, "Optimizing condition based maintenance," *Plant Engineering and Maintenance*, vol. 23, no. 6, pp. 57–71, 2001.

[21] Y. Liao, G. Lang, and L. Qu, "Precession trend analysis and balancing strategy for rotors with multi-fault," *Journal of Mechanical Engineering*, vol. 45, no. 8, pp. 45–51, 2009.

[22] L. Dieulle, C. Bérenguer, A. Grall, and M. Roussignol, "Sequential condition-based maintenance scheduling for a deteriorating system," *European Journal of Operational Research*, vol. 150, no. 2, pp. 451–461, 2003.

[23] M. Kearney, J. Marshall, and B. Newman, "Comparison of reliability enhancement tests for electronic equipment," in *Proceedings of the Reliability and Maintainability Symposium*, pp. 435–440, Singapore, January 2003.

[24] R. Ahmad and S. Kamaruddin, "An overview of time-based and condition-based maintenance in industrial application," *Computers and Industrial Engineering*, vol. 63, no. 1, pp. 135–149, 2012.

[25] R. D. Brillhart, D. L. Hunt, and H. Chimerine, "Multiple input excitation methods for aircraft ground vibration testing," *Sound and Vibration*, vol. 27, no. 1, pp. 77–85, 1993.

[26] Y. Umeda, T. Takeda Tomiyama et al., "Function behavior and structure," *Application of Artificial Intelligence in Engineering*, vol. 10, no. 4, pp. 177–193, 1990.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 283718, 9 pages
http://dx.doi.org/10.1155/2014/283718

Research Article

Aero-Engine Fault Diagnosis Using Improved Local Discriminant Bases and Support Vector Machine

Jianwei Cui,[1] Mengxiao Shan,[1] Ruqiang Yan,[1] and Yahui Wu[2]

[1] *School of Instrument Science and Engineering, Southeast University, Nanjing, Jiangsu 210096, China*
[2] *Changcheng Institute of Metrology and Measurement, Aviation Industry Corporation of China, Key Laboratory of Science and Technology on Metrology & Calibration, Beijing 100095, China*

Correspondence should be addressed to Ruqiang Yan; ruqiang@seu.edu.cn

Received 30 April 2014; Accepted 10 June 2014; Published 26 June 2014

Academic Editor: Weihua Li

This paper presents an effective approach for aero-engine fault diagnosis with focus on rub-impact, through combination of improved local discriminant bases (LDB) with support vector machine (SVM). The improved LDB algorithm, using both the normalized energy difference and the relative entropy as quantification measures, is applied to choose the optimal set of orthogonal subspaces for wavelet packet transform- (WPT-) based signal decomposition. Then two optimal sets of orthogonal subspaces have been obtained and the energy features extracted from those subspaces appearing in both sets will be selected as input to a SVM classifier to diagnose aero-engine faults. Experiment studies conducted on an aero-engine rub-impact test system have verified the effectiveness of the proposed approach for classifying working conditions of aero-engines.

1. Introduction

Aero-engine is one of the key components in an airplane and its reliability directly affects the flight safety of the airplane. However, in order to maintain good performance under high speed running condition, the clearance between rotor and stator in the aero-engine is getting smaller and smaller. This increases the possibility of rub-impact [1], which will generate unexpected vibrations, making the aero-engine not functioning well and even causing catastrophic consequences. Therefore, identifying the rub-impact fault in the aero-engine at its early stage is of great significance to both research and industrial communities.

As it is known that the rub-impact fault information is often carried by the weak transient vibrations, which are mixed together with other vibration sources, as a result, it is difficult to observe the fault symptom directly from the measured signals. With the development of modern signal processing, some advanced technologies, such as wavelet transform and Hilbert-Huang transform, have been utilized as viable tools for extracting fault-related features from vibration signals. As a classical time-frequency analysis method with solid mathematical foundation, wavelet transform in both continuous and discrete forms has been widely used for fault diagnosis [2–7]. As an extension of the discrete wavelet transform (DWT), the WPT has also been successfully applied to the field of fault diagnosis. For example, Boškoski and Juričić [8] proposed a novel approach for the diagnosis of gearboxes in presumably nonstationary and unknown operating conditions by making use of information indices based on the Renyi entropy derived from coefficients of the WPT of measured vibration records. Shen et al. [9] extracted statistical parameters from the signals obtained via the WPT at different decomposition depths and proposed a support vector regressive- (SVR-) based generic multiclass solver to identify the different fault patterns of rotating machinery. Keskes et al. [10] utilized stationary WPT for feature extraction under lower sampling rate to detect broken-rotor-bar and used the multiclass SVM to automatically recognize the faults. They utilized WPT to decompose multiclass signals into a library of time-frequency subspaces and calculated the wavelet packet energy in each subspace to produce a feature vector in each signal for classification [11].

Among these researches, most of researchers use the wavelet packet coefficients in the last decomposition level to extract the defect features of signals. It should be noted that the WPT has various wavelet packet subbands; thus there are multiple ways ($> 2^L$) to analyze a signal using a L-level decomposition. This implies that the subbands in the last decomposition level may not best reflect the signal feature and makes it necessary to optimize the decomposition process and improve its effectiveness. A widely applied criterion for optimal WPT-based signal decomposition is the Shannon entropy, which can be used to identify orthogonal subspaces with high-energy concentration that correlate with transients of interest by search for the minimum Shannon entropy [12]. But this criterion is mainly for signal representation. For classification problem, it is better to find optimal set of orthogonal subspaces that can yield high discriminant information for differentiating various classes as much as possible. In this study, local discriminant bases (LDB) algorithm has been employed to solve this problem. It selects the optimal set of orthogonal subspaces that can provide maximum dissimilarity information among different classes [13, 14]. Up to date, LDB has been applied to deal with real-world classification problems in the areas of audio signal analysis [15, 16], physiological signal classification [17, 18], and vibration data processing [13, 19]. From these applications, it can be seen that the results of LDB algorithm for a given dataset are driven by the nature of the dataset and the dissimilarity measures. At present, various dissimilarity measures, such as Euclidean distance, symmetric relative entropy, relative entropy, energy difference, correlation index, and nonstationarity, have been successfully utilized in many cases. In fact, accuracy of the classification results is highly influenced by the extent of class separation in feature space generated by the chosen dissimilarity measure and most researchers mainly use a single discriminant measure for the optimal subspace selection.

Motivated by these research efforts, an integrated approach that combines improved LDB algorithm with SVM is investigated for area-engine fault diagnosis in this study. The improved LDB utilizes two outstanding dissimilarity measures to choose the optimal set of orthogonal subspaces derived from WPT, and SVM obtains input from energy features derived from the optimal wavelet packet subspaces to classify working conditions of the aero-engine. This paper is organized as follows. Section 2 introduces the principle of the WPT; then the improved LDB algorithm is illustrated in Section 3; subsequently, Section 4 presents a multiclass classification method based on SVM. After that, the scheme for fault diagnosis using improved LDB and SVM is described and experiment study is conducted on an aero-engine rub-impact device to verify the effectiveness of the proposed method in Section 5. Finally, conclusions are drawn in Section 6.

2. Brief Introduction of WPT

WPT is an extension of DWT and can be obtained by a generalization of the fast pyramidal algorithm [20]. Mathematically, a wavelet packet consists of a set of linearly combined wavelet functions, which are generated using the following recursive relationships:

$$\begin{aligned} \psi^{2k}(t) &= \sqrt{2}\sum_n h(n)\,\psi^k(2t-n), \\ \psi^{2k+1}(t) &= \sqrt{2}\sum_n g(n)\,\psi^k(2t-n), \end{aligned} \tag{1}$$

where $\psi^0(t) = \phi(t)$ is the scaling function and $\psi^1(t) = \psi(t)$ is the wavelet function. The symbols $h(n)$ and $g(n)$ represent coefficients of a pair of quadrature mirror filters (QMF) associated with the scaling function and the wavelet function. Furthermore, $h(n)$ and $g(n)$ are related to each other by $g(n) = (-1)^n h(1-n)$. Using the QMF, a time-domain signal $\alpha(t)$ can be decomposed recursively as

$$\begin{aligned} \alpha_{j+1}^{2k}(t) &= \sum_m h(m-2n)\,\alpha_j^k(t), \\ \alpha_{j+1}^{2k+1}(t) &= \sum_m g(m-2n)\,\alpha_j^k(t), \end{aligned} \tag{2}$$

where $\alpha_j^k(t)$ denotes the wavelet packet coefficients at the jth level and kth subband. The symbol m represents the number of the wavelet coefficients at the kth subband within the level j. Using this equation, each detailed coefficient vector and approximation coefficient vector can be both decomposed into two parts and then a signal contained in $\Omega_{0,0}$ space can be decomposed into 2^j wavelet packet nodes (denoted as subspace $\Omega_{j,k}$) with the form of a full binary tree as shown in Figure 1. Each subspace $\Omega_{j,k}$ can be spanned by a series of base vectors $\{\alpha_{j,k,m}\}_{m=0}^{m=2^{N-j}-1}$, where 2^N corresponds to the length of the signal. Then a signal x_i can be represented by a set of coefficients as

$$x_i = \sum_{j,k,m} \left[\alpha_{j,k,m}\right]_i \cdot \omega_{j,k,m}. \tag{3}$$

Through the 3-level decomposition process as shown in Figure 1, it can be seen that the WPT has various styles for the selection of orthogonal subspaces, such as $\{\Omega_{3,0}, \Omega_{3,1}, \Omega_{3,2}, \Omega_{3,3}, \Omega_{3,4}, \Omega_{3,5}, \Omega_{3,6}, \Omega_{3,7},\}$, $\{\Omega_{2,0}, \Omega_{3,2}, \Omega_{3,3}, \Omega_{3,4}, \Omega_{3,5}, \Omega_{3,6}, \Omega_{3,7},\}$, or $\{\Omega_{2,0}, \Omega_{2,1}, \Omega_{2,2}, \Omega_{3,6}, \Omega_{3,7}\}$. Therefore, the optimal selection of orthogonal subspace set needs to be investigated.

3. Improved LDB Algorithm

The LDB algorithm is a pruning algorithm which identifies the subspaces that exhibit high discrimination between signal classes by using a given dissimilarity measure [21]. LDB selects an orthogonal basis from a dictionary of bases in a wavelet packet to distinguish different classes in a given set of data belonging to several classes and is used to select the optimal set of complete orthogonal subspaces derived from the WPT.

Suppose that $A_{j,k}$ represents the desired local discriminant basis restricted to the span of $B_{j,k}$, which is a set of basis vectors at (j,k) node. Then, for a given dataset consisting of L classes of signals $\{\{x_i^{(l)}\}_{i=1}^{N_l}\}_{l=1}^{L}$ with N_l being the total number

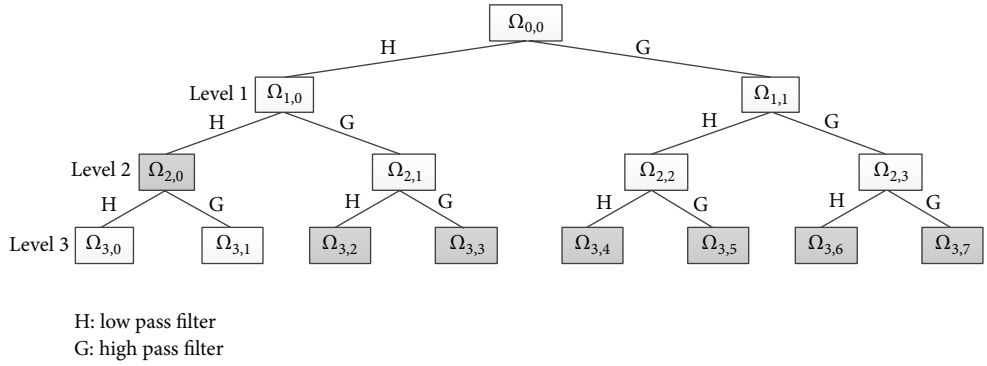


Figure 1: A 3-level decomposition tree in a wavelet packet.

of signals in class l, the traditional LDB algorithm with an additive dissimilarity measure D can then be summarized as follows.

(1) The WPT is used to decompose the signals contained in the training dataset.

(2) The time-frequency energy maps C_l for $l = 1, \ldots, L$ on the wavelet packet coefficients according to (3) and (4) are constructed:

$$C_l(j,k,m) \equiv \frac{\sum_{i=1}^{N_l} \left(\omega_{j,k,m}^T x_i^{(l)}\right)^2}{\sum_{i=1}^{N_l} \left\|x_i^{(l)}\right\|^2}. \tag{4}$$

(3) Assume $A_{J,k} = B_{J,k}$ and set $\Delta_{J,k} = D(\{C_l(J,k,\cdot)\})_{l=1}^{L}$, where the array containing the dissimilarity measure of the node (j,k), for $k = 0, \ldots, 2^{J-1}$, the best subspaces $A_{j,k}$, can be obtained through the following condition.

If $\Delta_{j,k} \geq \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$, namely, the dissimilarity measure of the parent node is greater than those of cumulative dissimilarity measure of the children nodes, then $A_{j,k} = B_{j,k}$.

Else $A_{j,k} = A_{j+1,2k} \oplus A_{j+1,2k+1}$ and set $\Delta_{j,k} = \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$.

(4) After a complete set of orthogonal subspaces are found in the decomposition results, their corresponding basis functions are ranked from higher to lower according to their discrimination power, and the t (much less than n) most discriminant basis functions can be used for constructing classifiers.

From the algorithm above, it should be noted that the optimal choice of LDB subspaces for a given dataset is significantly affected by the dissimilarity measures used to distinguish among classes. The dissimilarity measure indirectly controls the classification accuracy achieved. In order to obtain good classification results, a significant dissimilarity measure, which is capable of discriminating among different classes as much as possible, should be studied. However, when dealing with complex datasets such as the vibration signals of aero-engines, using a single dissimilarity measure for the optimal subspace selection may not be able to capture all the characteristic information of its class while using multiple dissimilarity measures provides additional feature dimensions for classification. Hence, instead of using a single dissimilarity measure, a combination of two dissimilarity measures (D_1 and D_2) with varying complexity is studied to select the LDB with different characteristics to achieve high classification accuracies in the presented approach.

The first dissimilarity measure D_1 is defined as the difference in the normalized energy between the corresponding wavelet packet nodes of the training signals from different classes. The normalized energy difference D_1 is given by

$$D_1^{(1,2)} = \left|E_{j,k}^1 - E_{j,k}^2\right|, \tag{5}$$

where $E_{j,k}^1$ and $E_{j,k}^2$ are the normalized energy of the corresponding wavelet packet nodes (j,k), which can be calculated by

$$E_{j,k} = \frac{\sum_{m=0}^{m=2^{n_0-j}-1} \left(\alpha_{j,k,m}^2\right)}{E_{x_i}}, \tag{6}$$

where $j = 0, 1, \ldots, J$, $k = 0, 1, \ldots, 2^j - 1$, $n_0 = \log_2 n \geq J$ (n is the signal size and n_0 is the maximum level of signal decomposition). In addition, $\alpha_{j,k,m}$ is the wavelet packet coefficient of the corresponding nodes (j,k) at position (m) and E_{x_i} represents the total energy of the vibration signals.

The second dissimilarity measure D_2 calculates the distribution difference of two classes at the wavelet node (j,k), which is described as the relative entropy and expressed as

$$D_2^{(1,2)} = \sum_{i=1}^{n} p_i^{(1)} \log \frac{p_i^{(1)}}{p_i^{(2)}}, \tag{7}$$

where $n = 2^{n_0-j} - 1$, $\sum_i p_i^{(1)} = \sum_i p_i^{(2)} = 1$, and $p_m(j,k) = \alpha_{j,k,m}^2 / \sum_{i=1}^{n} |\alpha_{j,k,i}|^2$ stands for the energy proportion of some wavelet coefficient $\alpha_{j,k,m}$ making up the total energy of the wavelet node (j,k).

It can be seen in (5) and (7) that D_1 and D_2 are always nonnegative and will be zero if distributions of $E_{j,k}$ or p from two classes are the same. Furthermore, the further the two

distributions are, the higher the dissimilarity measures D_1 and D_2 will be.

Similarly, for multiple class ($L > 2$) problems, the normalized energy difference and the relative entropy can be expressed as

$$D_1 = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} D_1^{(i,j)} \qquad D_2 = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} D_2^{(i,j)}. \tag{8}$$

Based on the normalized energy difference and relative entropy, the improved LDB selection process in searching the optimal wavelet packet subspaces is shown in Figure 2 and described below.

The vibration signals are first decomposed by the WPT. Then the normalized energy difference and relative entropy of each subspace are calculated among classes using (5) and (7). After that, the wavelet packet tree is pruned from bottom to top according to the following rules: if the discriminative measure of the parent node is larger than that of the cumulative discriminative measure of the children nodes, the parent node is kept and the children nodes need to be deleted; otherwise, the children nodes need to be kept and the dissimilarity measure of the parent node should be set as the sum of the dissimilarity measure of the children nodes. At the end of this iterative process, the tree structure contains only those terminal nodes, which contribute to maximizing the distance among different classes. Since we utilized two dissimilarity measures, two optimal local discriminant bases are obtained at last. As D_1 is expected to reveal the energy concentration locations on the time-frequency plane for different types of vibration signals while D_2 describes the degree of separation between different distribution series, the nodes that exist in both sets possess high discriminatory values among all the classes for both of the given dissimilarity measures and can be used to form feature vectors.

In this study, the energy feature of each subspace is investigated for constructing the feature vector. The energy of each subspace is defined as

$$E_t = \sum_{i=1}^{M} \alpha_t(i)^2, \tag{9}$$

where M is the number of the wavelet packet coefficients in each subspace and $\alpha_t(i)$ is the wavelet packet coefficient. Then, for the t chosen subspaces that exist in both sets in LDB, a feature vector can be constructed from all the subspaces as

$$F = [E_1, E_2, \ldots, E_t]. \tag{10}$$

The vector F will be selected as input to a classifier for identifying aero-engine working conditions.

4. Multiclass SVM Classifier

SVM is a linear learning method that finds an optimal hyperplane to separate two classes. As a supervised classification approach, SVM seeks to maximize the distance to the closest training point from either class in order to achieve better classification performance on test data [22]. Due to the small-sample characteristic of the SVM, it is suitable to distinguish different classes with a small number of data. As the training data is often limited in real-time fault diagnosis, SVM is utilized as a classifier in this study to diagnose different aero-engine working conditions. However, the SVM cannot be directly applied to the multiclassification problems since traditional SVM is designed to deal with the two-class problem. For multiclass problems, SVM can solve this dilemma through the combination of two-class problems.

The crucial widely used multiclass SVM (MSVM) strategy is the one-against-all (OAA) strategy and one-against-one (OAO) strategy. OAA is the simplest MSVM strategies. It involves k binary SVM classifiers, one for each class. Each binary SVM is trained to separate one class from the rest. The winning class is the one that corresponds to the SVM with the highest output. OAO involves $k(k-1)/2$ binary SVM classifiers. Each classifier is trained to separate each pair of classes. The advantage of OAA is the fastness of classification; therefore, a multiclass classification method based on OAA strategy is used in this study.

The multiclass classification method can be clearly described in Figure 3: for K-class sample training, $K-1$ SVMs are trained and the first samples are seen as positive samples while the other $K-1$ classes are viewed as negative samples to train the SVM1; then the first samples are removed, and the same process will repeat until the $(K-1)$th classifier is designed. During the test process, the samples are treated as input to the first classifier and the test will be over only if the output is "1," which means that the sample class is the corresponding category of the classifier; otherwise, the samples will be sent to the next classifier until the test samples are distinguished.

5. Fault Diagnosis Scheme with Experimental Verification

Following the knowledge as explained in previous sections, the proposed aero-engine fault diagnosis approach is depicted in Figure 4. It includes two parts: training and testing parts. For training, vibration signals from each of the working conditions are decomposed into wavelet packet trees with a selected wavelet function. After that, the corresponding nodes of the trees are compared using a set of dissimilarity measures to identify the nodes that exhibit high discriminative values among various aero-engine working conditions. After selecting the significant LDB nodes, a new wavelet packet tree is constructed, and all the signals are then decomposed using this new wavelet packet tree. Features are finally extracted from the LDB nodes to train a multiclass SVM classifier. For testing, energy features which are extracted from those selected LDB nodes are input to the trained multiclass SVM classifier for working condition identification.

In order to verify the effectiveness of the proposed aero-engine fault diagnosis approach, an experimental study was carried out on a twin-shaft aero-engine test system. The vibration signals were acquired at 64 kHz sampling rate by a velocity sensor, which was mounted on the outside of

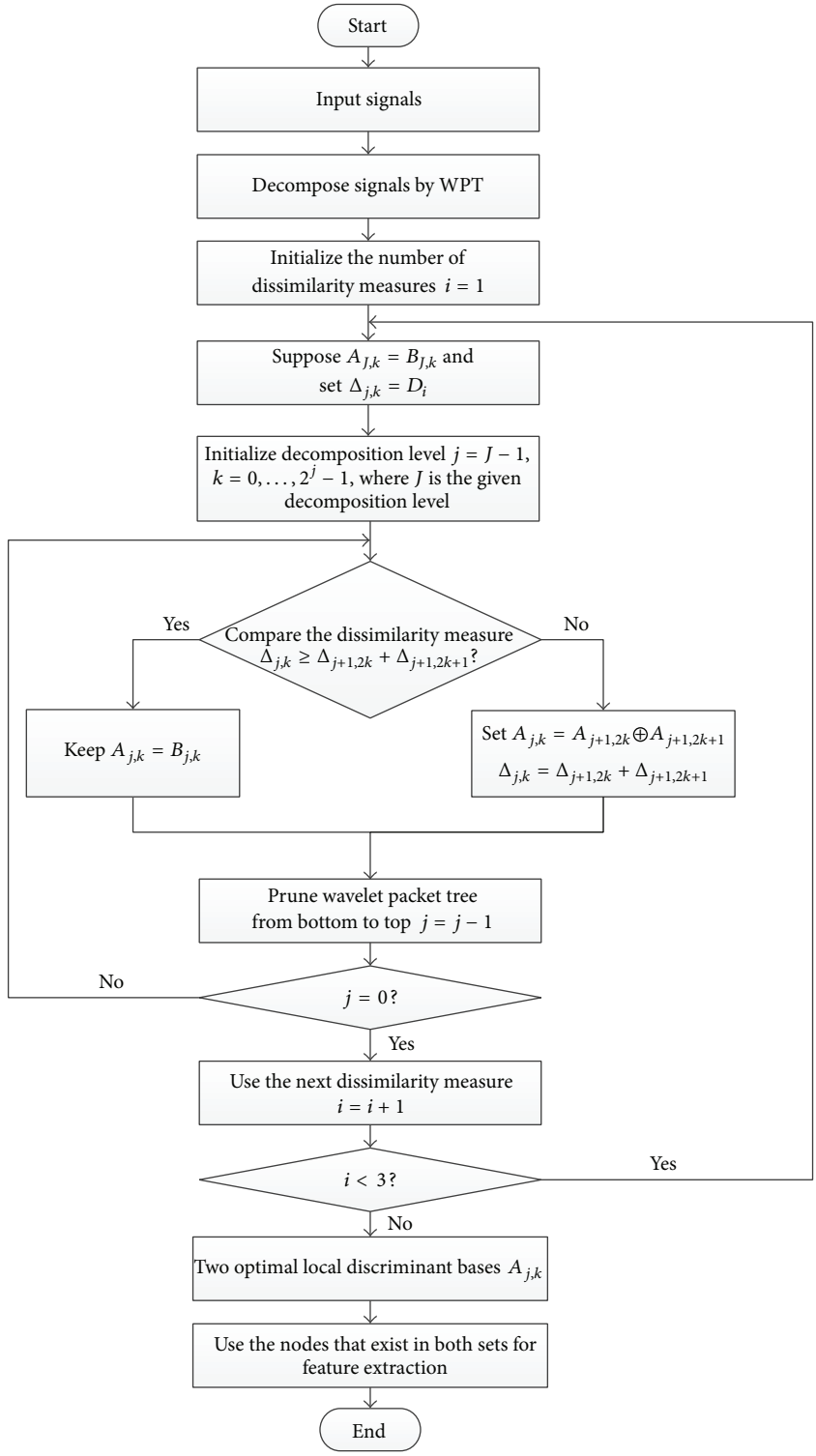


Figure 2: Flow chart of the improved LDB algorithm.

the aero-engine casing. Due to the complex structure of the aero-engine, the signals often contain vibrations generated by low pressure shaft, high pressure shaft, and the transmission system, causing nonstationarity. Three different working conditions, that is, faultless, rub-impact fault, and unbalance fault, were considered in this study. Figure 5 shows waveforms of the sampled signals.

The proposed approach is then used to process the vibration signals. It should be noted that an appropriate wavelet function should be chosen for WPT, as it will affect

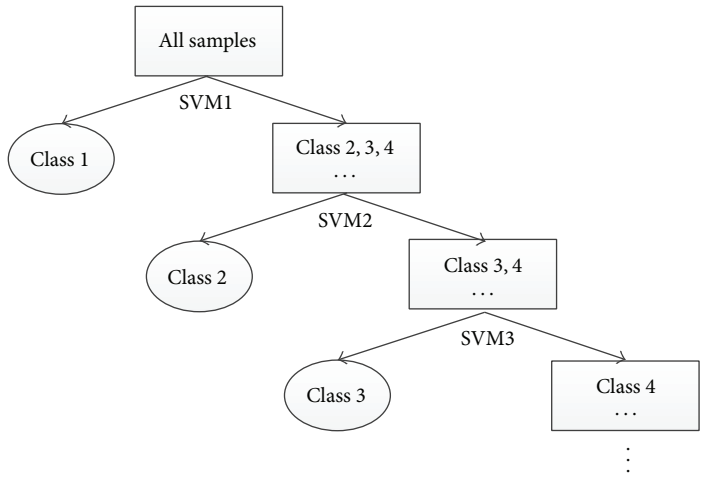


Figure 3: The multiclass SVM classifier.

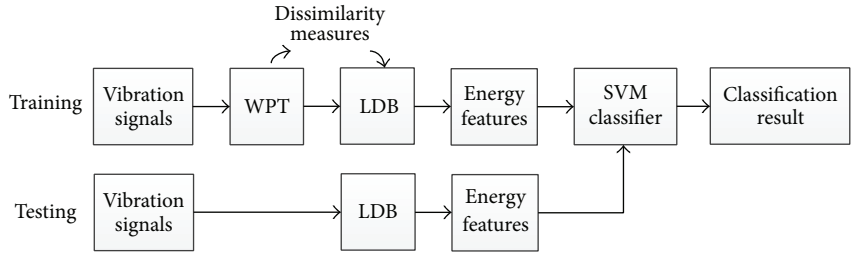


Figure 4: Block diagram of the aero-engine fault diagnosis scheme.

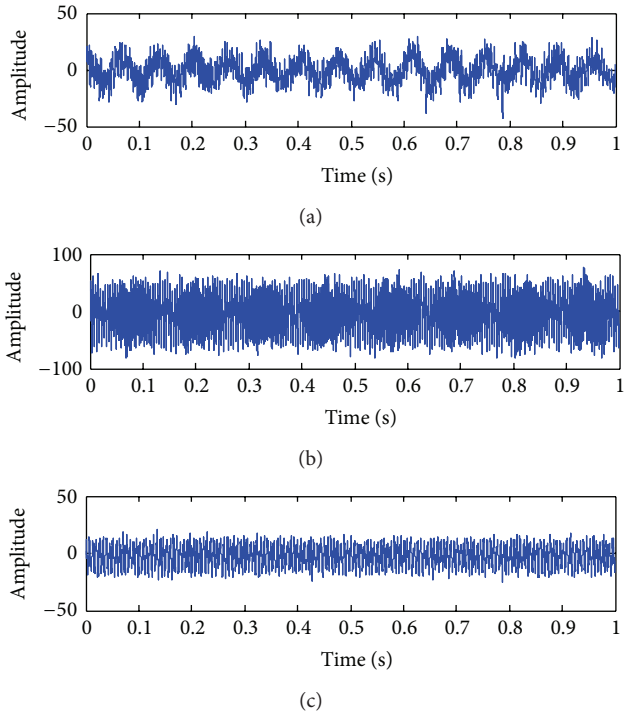


Figure 5: Aero-engine vibration signals of three different working conditions: (a) faultless, (b) rub-impact, and (c) unbalance.

the decomposition performance. In this study, a mutual information criterion is used for guiding the selection of wavelet function [23]. In information theory, mutual information is usually used to measure the degree of similarity between two groups of data sequence. The greater the mutual information is, the more similar the two groups of data sequence will be. Such a relationship is applicable to wavelet function selection for aero-engine fault diagnosis by taking the vibration signal and wavelet packet coefficients as data sequences X and Y, respectively. By comparison, the wavelet function that maximizes the mutual information between the vibration signal and the reconstruction signal represents the most appropriate wavelet for rub-impact vibration extraction. Based on this criterion, a total of 30 candidate wavelet functions (e.g., Haar, Db2, Db4, Coif1, Coif2, Bior1.3, Bior5.5, etc.) are evaluated, and the Bior5.5 wavelet is considered as the most appropriate wavelet function to process the rub-impact signals. After that the aero-engine vibration signals are processed using the selected wavelet function for a 4-level decomposition and the improved LDB is utilized to select the optimal subspaces derived from the decomposition results. Figure 6 shows the selected wavelet packet nodes (marked with black block) that contain the best discriminant information to classify different working conditions using the dissimilarity measures D_1 (normalized energy difference) and D_2 (relative entropy), respectively. These blocks in each figure represent

Table 1: Results of the experimental study.

Target classes	Sample size	Recognition results			Recognition rate [%]	Total [%]
		Faultless	Rub-impact	Unbalance		
Faultless	20	20	0	0	100	100.00
Rub-impact	20	0	20	0	100	
Unbalance	20	0	0	20	100	

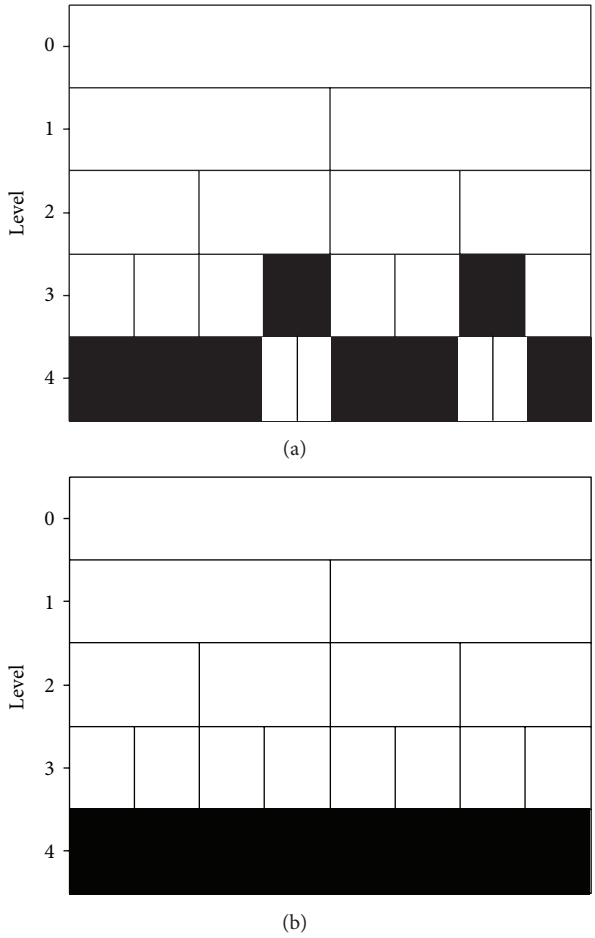


Figure 6: (a) Selected wavelet packet nodes (dissimilarity measure D_1, bior5.5 wavelet). (b) Selected wavelet packet nodes (dissimilarity measure D_2, bior5.5 wavelet).

Table 2: Classification performance using different dissimilarity measures.

Dissimilarity measures	Recognition rate [%]
D_1	95.00
D_2	91.67
D_1 and D_2	100.00

the complete information of the signal with the capability of differentiating various aero-engine working conditions, as manifested by the nature of LDB algorithm.

For two dissimilarity measures (D_1 and D_2), altogether 30 LDB nodes as shown in Figure 6 are identified. Some of the nodes are selected by both dissimilarity measures. These nodes that exist in both of the LDB trees demonstrate relatively high discriminatory behavior among the combinations of all working conditions for both of the given dissimilarity measures. In other words, these nodes demonstrate high statistical distance among all working conditions for both of the given dissimilarity measures. Therefore, the LDBs appearing in both of the LDB trees are used to extract features.

Table 3: Classification performance using different classifiers.

Classifier types	Recognition rate [%]
SVM classifier	100.00
Bayes classifier	98.33
HMM classifier	78.33
BP NN classifier	98.33

Generally, the basis vector coefficients from each of the selected LDB nodes can be directly used as features. However, considering that more features may not necessarily increase the performance of a given classifier, the energy content of the selected LDBs calculated by (9) is extracted as features. In this study, the energy values of the 12 LDB nodes that exist in both of the LDB trees (40 groups of training signals and 20 groups of testing signals, each containing 1,024 data points) are input to the multiclass SVM classifier for characterizing aero-engine working conditions.

Table 1 lists the classification results of this experimental study. It can be seen that the SVM classifier results in much high classification accuracies scoring 100%, which indicates that the developed approach is suitable for aero-engine fault diagnosis.

For the purpose of performance comparison, the single dissimilarity measure is also used to select the LDB nodes and the corresponding energy features are used as input to the SVM classifier. The classification results are shown in Table 2, which indicates that the diagnosis approach using multiple dissimilarity measures can achieve better classification performance than that using single dissimilarity measure.

The effect of different classifiers, such as the Bayes classifier, hidden Markov model (HMM) classifier, and back-propagation (BP) neural network (NN) classifier, on the classification performance, is also studied. As it is shown in Table 3, the SVM classifier performs the best; this is contributed by its good ability of dealing with small size of samples.

In addition, the effect of wavelet functions on the classification performance is investigated in this study. Three difference wavelet functions, including Haar wavelet, Db2 wavelet, and Bior5.5 wavelet, are used to process the aero-engine

Table 4: Classification performance using different wavelet functions.

Wavelet function	Classification accuracy [%]			Recognition rate [%]
	Faultless	Rub-impact	Unbalance	
Haar	100	100	85	95.00
Db2	100	100	85	95.00
Bior5.5	100	100	100	100.00

vibration signals, and the final classification performance is shown in Table 4. It can be seen that the Bior5.5 wavelet function chosen by the quantitative mutual information measure leads to higher classification rate than the other two wavelet functions.

6. Conclusions

Based on the improved LDB and SVM, an integrated approach for aero-engine fault diagnosis has been developed. The results of experimental study conducted on an aero-engine test system indicate that the proposed approach has good ability to classify different aero-engine working conditions. Furthermore, the comparison study shows that the improved LDB algorithm can improve the classification accuracy, and an appropriate wavelet function provides better signal decomposition. Further study is being conducted for providing effective and efficient solutions on aero-engines condition monitoring and fault diagnosis.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work has been supported in part by the National Natural Science Foundation of China under Grant 51175080, the Aeronautical Science Foundation of China under Grant 20122269015, and the Science and Technology Support Program of Jiangsu Province under Grant BE2012740.

References

[1] Y. Wu, J. Xue, D. Zhang, and X. Li, "Research on aeroengine rub-impact fault analysis based on wavelet transform and the local binary patterns," in *Proceedings of the 8th International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR '10)*, pp. 421–426, Qingdao, China, July 2010.

[2] Z. K. Zhu, Z. He, A. Wang, and S. Wang, "Synchronous enhancement of periodic transients on polar diagram for machine fault diagnosis," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 7, no. 4, pp. 427–442, 2009.

[3] R. Yan and R. X. Gao, "Harmonic wavelet-based data filtering for enhanced machine defect identification," *Journal of Sound and Vibration*, vol. 329, no. 15, pp. 3203–3217, 2010.

[4] S. Wang, W. Huang, and Z. K. Zhu, "Transient modeling and parameter identification based on wavelet and correlation filtering for rotating machine fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 25, no. 4, pp. 1299–1320, 2011.

[5] J. Liu, "Shannon wavelet spectrum analysis on truncated vibration signals for machine incipient fault detection," *Measurement Science and Technology*, vol. 23, no. 5, Article ID 055604, 2012.

[6] R. Kumar and M. Singh, "Outer race defect width measurement in taper roller bearing using discrete wavelet transform of vibration signal," *Measurement*, vol. 46, no. 1, pp. 537–545, 2013.

[7] P. Li, F. Kong, Q. He, and Y. Liu, "Multiscale slope feature extraction for rotating machinery fault diagnosis using wavelet analysis," *Measurement*, vol. 46, no. 1, pp. 497–505, 2013.

[8] P. Boškoski and D. Juričić, "Fault detection of mechanical drives under variable operating conditions based on wavelet packet Rényi entropy signatures," *Mechanical Systems and Signal Processing*, vol. 31, pp. 369–381, 2012.

[9] C. Shen, D. Wang, F. Kong, and P. W. Tse, "Fault diagnosis of rotating machinery based on the statistical parameters of wavelet packet paving and a generic support vector regressive classifier," *Measurement*, vol. 46, no. 4, pp. 1551–1564, 2013.

[10] H. Keskes, A. Braham, and Z. Lachiri, "Broken rotor bar diagnosis in induction machines through stationary wavelet packet transform and multiclass wavelet SVM," *Electric Power Systems Research*, vol. 97, pp. 151–157, 2013.

[11] Q. He, "Vibration signal classification by wavelet packet energy flow manifold learning," *Journal of Sound and Vibration*, vol. 332, no. 7, pp. 1881–1894, 2013.

[12] R. Gao and R. Yan, "Non-stationary signal processing for bearing health monitoring," *International Journal of Manufacturing Research*, vol. 1, no. 1, 2006.

[13] Q. He, R. Yan, and R. X. Gao, "Wavelet packet base selection for gearbox defect severity classification," in *Proceedings of the Prognostics and Health Management Conference (PHM '10)*, pp. 1–5, Macau, China, January 2010.

[14] Y. Wu, M. Shan, Y. Qian, X. Li, and R. Yan, "Aeroengine rub-impact fault diagnosis based on wavelet packet transform and the local discriminate bases," *Applied Mechanics and Materials*, vol. 226–228, pp. 740–744, 2012.

[15] K. Umapathy, S. Krishnan, and R. K. Rao, "Audio signal feature extraction and classification using local discriminant bases," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1236–1246, 2007.

[16] P. T. Hosseini, F. Almasganj, and M. R. Darabad, "Pathological voice classifcation using local discriminant basis and genetic algorithm," in *Proceedings of the 16th Mediterranean Conference on Control and Automation (MED '08)*, pp. 872–876, Ajaccio, France, June 2008.

[17] A. R. Harris, K. Schwerdtfeger, and D. J. Strauss, "Optimized shift-invariant wavelet packet feature extraction for electroencephalographic evoked responses," in *Proceedings of the 30th*

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '08), pp. 2685–2688, Vancouver, Canada, August 2008.

[18] A. Taki, O. Pauly, S. K. Setarehdan, G. Unal, and N. Navab, "IVUS-based histology of atherosclerotic plaques: improving longitudinal resolution," in *Proceedings of the Medical Imaging: Ultrasonic Imaging, Tomography, and Therapy*, February 2010.

[19] Z. Zhuang and F. Li, "Statistical method for rotating machine fault diagnosis," in *Proceedings of the International Conference on Manufacturing Science and Technology*, 2011.

[20] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, Calif, USA, 1999.

[21] N. Saito and R. R. Coifman, "Local discriminant bases and their applications," *Journal of Mathematical Imaging and Vision*, vol. 5, no. 4, pp. 337–358, 1995.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, NY, USA, 2001.

[23] R. Gao and R. Yan, *Wavelets, Theory and Applications for Manufacturing*, Springer, New York, NY, USA, 2010.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 923424, 12 pages
http://dx.doi.org/10.1155/2014/923424

Research Article

Machine Fault Classification Based on Local Discriminant Bases and Locality Preserving Projections

Qingbo He,[1] Xiaoxi Ding,[1] and Yuanyuan Pan[2]

[1] *Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, Anhui 230026, China*
[2] *Anhui Vocational College of City Management, Hefei, Anhui 231635, China*

Correspondence should be addressed to Qingbo He; qbhe@ustc.edu.cn

Received 27 April 2014; Accepted 1 June 2014; Published 26 June 2014

Academic Editor: Weihua Li

Machine fault classification is an important task for intelligent identification of the health patterns for a mechanical system being monitored. Effective feature extraction of vibration data is very critical to reliable classification of machine faults with different types and severities. In this paper, a new method is proposed to acquire the sensitive features through a combination of local discriminant bases (LDB) and locality preserving projections (LPP). In the method, the LDB is employed to select the optimal wavelet packet (WP) nodes that exhibit high discrimination from a redundant WP library of wavelet packet transform (WPT). Considering that the obtained discriminatory features on these selected nodes characterize the class pattern in different sensitivity, the LPP is then applied to address mining inherent class pattern feature embedded in the raw features. The proposed feature extraction method combines the merits of LDB and LPP and extracts the inherent pattern structure embedded in the discriminatory feature values of samples in different classes. Therefore, the proposed feature not only considers the discriminatory features themselves but also considers the dynamic sensitive class pattern structure. The effectiveness of the proposed feature is verified by case studies on vibration data-based classification of bearing fault types and severities.

1. Introduction

Machine fault classification is an important task for intelligent identification of the condition patterns for the system being monitored. For a mechanical system, vibration monitoring is often employed to evaluate the system dynamics. A specific application considered in this paper is to monitor health condition of a machine or its components, such as bearings, for timely identifying possible faults, which is increasingly significant to reduce machine downtime and ensure high productivity. Once a fault happens in a machine, it makes sense to identify the fault type or the fault severity through vibration data analysis so that time and safety can be guaranteed. There are many reasons leading to machine failures. For instance, poor lubrication, acid corrosion, and plastic deformation could cause the bearing to work in an abnormal condition, respectively [1]. In addition, typical damages of the rolling bearing are located at outer raceway, inner raceway, or rolling element. To effectively monitor and recognize the machine condition, the major challenge is to extract reliable features from vibration data which are often disturbed by the environment noise. The traditional features, such as the time-domain features and the frequency-domain ones, are often applied to fault diagnosis [2–6]. However, the pattern of vibration signals demonstrates many nonlinear characteristics and the methods mentioned above cannot extract these nonlinear features effectively for classifying fault types and severities. Therefore, this study intends to find a good feature representation of the raw signals that yields higher discriminatory information.

Wavelet transform has the ability to well express the nonstationary signals and represent sensitive features with its multiresolution capability, which has achieved a great success in fault classification [7]. As one of the most widely used wavelet transform methods, the wavelet packet transform (WPT) is well-known for its orthogonal, complete, and local property [8]. WPT leads to a redundant binary tree of a signal with a set of time-frequency subspaces each of which is made

up of a wavelet packet (WP) base vector. The whole subspaces are called a WP library. As we know, different WP bases give rise to different representation of a given signal. Thus, it is important to select optimal WP bases out of the whole WP library for enhanced signal analysis or classification. For classification, the main objective is to find an optimal set of WP nodes that yield high discriminatory information for discriminating different classes as much as possible. This can be realized by the local discriminant bases (LDB) [9]. The algorithm identifies optimal LDBs with high discriminatory information by using a dissimilarity measure on the given dataset. Many related works have been conducted in the last two decades to demonstrate the effectiveness of the LDB to achieve a good classification through selecting the optimal WP bases among various redundant WP subspaces [9–18]. Although the discriminatory features can be obtained by selecting WP nodes via the LDB, different node displays different sensitivity in characterizing class information. In the machine learning-based classification approach, the classification accuracy will mainly depend on the sensitive features. Current methods mainly employ the most sensitive bases for classification. We focus on another approach, which is using the dimensionality reduction techniques to mine more sensitive features in the whole set of discriminatory features by LDB. Therefore, in this study, one challenge is how to extract the most useful and sensitive information hidden in high-dimensional data based on the selected WP nodes.

In the past few decades, many useful dimensionality reduction techniques have been employed for fault diagnosis and classification [2, 3, 19–27]. These techniques can be simply divided into two types: linear and nonlinear approaches. Linear dimensionality reduction aims to find a set of low-dimensional bases from high-dimensional data through the linear transformation. Two of the well-known linear learning methods are principal component analysis (PCA) [19, 20] and linear discriminant analysis (LDA) [21, 22]. The other type, nonlinear dimensionality reduction, searches for nonlinear structure hidden in high-dimensional data. There are two main nonlinear approaches including kernel-based techniques [2, 23, 24] and manifold learning techniques [25–27]. Manifold learning pursuits the goal to embed data that originally lies in a high-dimensional space in a lower dimensional space while preserving local characteristic properties, for example, local geometric property (Isomap [28]), local embedding structure (LLE [29]), local adjacency relations (LE [30]), and local tangent space information (LTSA [31]). Although these nonlinear manifold learning methods have been effectively developed to machine fault classification, they need heavy computation cost and are complex to be extended for fault classification of a new data [25–27, 32]. He and Niyogi [33] proposed a new linear model, locality preserving projections (LPP), which can reveal the nonlinear manifold structure embedded in the dataset with a kernel that maintains the local information. LPP is provided with the remarkable superiority that it can form an explicit map to the manifold learning algorithm, which is linear and easily operational. Some works have indicated that LPP is beneficial to feature extraction in machine fault classification [3, 24]. Hence, the LPP is employed in this study to extract the sensitive information hidden in the raw feature data from the selected WP nodes.

In this paper, based on energy features of the nodes selected by LDB algorithm from the WP library, a new effective feature is proposed to mine the nonlinear pattern information by LPP in the case of bearing fault classification. The proposed feature intends to overcome the weakness of the discriminatory WP nodes for characterizing the fault pattern in different sensitivity. Specifically, vibration signals from the bearings with different fault types and severities are firstly decomposed into the WP library and the LDB is then applied to identify the optimal WP subspaces that supply maximum dissimilarity information among them. After that, the root energy of the selected nodes constitutes a raw feature set. Due to the redundant property of the features in representing the fault pattern, some important sensitive information may be submerged among them. Therefore, the LPP is employed to extract the nonlinear sensitive pattern information embedded in the dataset. These sensitive features are finally chosen as inputs to a diagnostic classifier for characterizing bearing types and severities.

The rest of this paper is organized as follows. Section 2 describes the theoretical background and major principle of the proposed feature extraction method that combines the LDB and the LPP. In Section 3, experimental results on bearing fault classification are used to verify the effectiveness of the proposed method as compared to other traditional feature extraction methods. Finally, conclusions are provided in Section 4.

2. Theoretical Background

2.1. WPT for Signal Decomposition. The WPT is an excellent signal decomposition tool with well-known properties of being orthogonal, complete, and local [8]. In operation, the WPT utilizes a series of low-pass and high-pass filters to filter a signal being analyzed recursively. Through this way, a signal $x(t)$ can be decomposed into a set of WP nodes with the form of a full binary tree by the WPT. Each node possesses a specific time-frequency subspace. Let $\Omega_{0,0}$ denote a vector space $\mathbb{R}^n$ corresponding to the node 0 of the parent tree. Then at each level the vector space is split into two mutually orthogonal subspaces by a pair of low-pass and high-pass filters. The split process can be given by

$$\Omega_{j,k} = \Omega_{j+1,2k} \oplus \Omega_{j+1,2k+1}, \tag{1}$$

where j indicates the level of the tree and k represents the node index in level j with $k = 0, \ldots, 2^j - 1$. This process is repeated until level J, giving rise to 2^J mutually orthogonal subspaces.

Each subspace $\Omega_{j,k}$ is spanned by a set of base vectors $\{w_{j,k,m}\}_{m=0}^{m=2^{n_0-j}-1}$, where $n_0 = \log_2 N \geq J$ (N is signal length and n_0 is the maximum level of signal decomposition). The vector $w_{j,k,m}$ represents the WP base function indexed by the triplet (j, k, m) representing scale, frequency band (oscillation), and time position, respectively.

The WP coefficients of signal $x(t)$ can be calculated in the inner product of the signal with every WP function as follows:

$$\alpha_{j,k,m} = \left\langle x, w_{j,k,m} \right\rangle = \int_{-\infty}^{\infty} x(t)\, w_{j,k,m}(t)\, dt, \tag{2}$$

where $\alpha_{j,k,m}$ denotes the kth set of WP coefficients at the jth scale parameter and m is the translation parameter. In other words, the signal $x(t)$ is decomposed into 2^J subspaces with coefficients $\{\alpha_{j,k,m}\}_{m=0}^{m=2^{n_0-j}-1}$ in each subspace.

The signal $x(t)$ can then be expressed as

$$x(t) = \sum_{j,k,m} \left[\alpha_{j,k,m}\right]_i \cdot w_{j,k,m}, \tag{3}$$

where, in the index (j,k,m), (j,k) corresponds to the terminal (leaf) nodes and $\alpha_{j,k,m}$ are the base vector coefficients at position (m).

2.2. LDB for WP Selection. The LDB is a pruning algorithm that identifies the subspaces and their bases that exhibit high discrimination between signal classes using a given dissimilarity measure [9]. The optimal selection of LDB subspaces for a given dataset is driven by the nature of the dataset and the dissimilarity measure. Dissimilarity measure is designed to evaluate the "statistical distances" among different classes for each WP node. Numerous dissimilarity measures have been developed so far, such as relative entropy, energy difference, correlation index, and nonstationarity. In this paper, relative entropy is investigated as the dissimilarity measure in identifying optimal WP subspaces.

The LDB algorithm is used to identify the WP nodes that exhibit high discrimination, indicated by large statistical distance between classes. A set of training signals for all L classes are decomposed into full binary WP trees of order J. Let each of the signals in the training set be denoted by x_i^l, where the index i and l correspond to the ith training signal in the lth class. The WP tree is pruned by the LDB algorithm in such a way that a node is split if the cumulative discriminative measure of the children nodes is greater than that of the parent node. In other words, a node is split only if the children nodes have better discriminative power than that of the parent node. As a result, the process will end with a subset of terminal WP nodes that contribute to maximizing the statistical distance between different classes.

Mathematically, the LDB selection process is described as follows. Suppose that $A_{j,k}$ represents the desired local discriminant base restricted to the span of $B_{j,k}$, which is a set of base vectors at (j,k) node, and $\Delta_{j,k}$ is the array containing the discriminant measure of the same node.

LDB Algorithm. A training dataset consisting of L class of signals $\{\{x_i^{(l)}\}_{i=1}^{N_l}\}_{l=1}^{L}$ with N_l being the total number of training signals in class l is given.

Step 1. Choose a time-frequency decomposition method, such as the WPT, to decompose the signals contained in the training dataset.

Step 2. Construct time-frequency energy maps C_l for $l = 1, \ldots, L$ on the WP coefficients. Here, C_l is calculated by accumulating the squares of expansion coefficients of the signals at each position followed by a normalization with respect to the total energy of all the training signals belonging to class l as follows:

$$C_l(j,k,m) \equiv \frac{\sum_{i=1}^{N_l} \left(w_{j,k,m}^T x_i^{(l)}\right)^2}{\sum_{i=1}^{N_l} \left\| x_i^{(l)} \right\|^2}, \tag{4}$$

where $j = 0, 1, \ldots, J$, $k = 0, 1, \ldots, 2^j - 1$, $m = 0, 1, \ldots, 2^{n_0-j} - 1$.

Step 3. Set $A_{J,k} = B_{J,k}$, where $B_{J,k}$ is the base set spanning subspace of $\Omega_{J,k}$ node (J,k) and then evaluate $\Delta_{J,k} = D(\{C_l(J,k,\cdot)\}_{l=1}^{L})$ for $k = 0, \ldots, 2^J - 1$. Let $p^{(l)} = C_l(J,k,\cdot)$; for multiple class problems, the dissimilarity measure based on relative entropy is expressed as

$$D\left(\left\{p^{(l)}\right\}_{l=1}^{L}\right) = \sum_{a=1}^{L-1} \sum_{b=a+1}^{L} \sum_{i=1}^{N_l} p_i^{(a)} \log \frac{p_i^{(a)}}{p_i^{(b)}}. \tag{5}$$

Step 4. Determine the best subspace $A_{j,k}$ for $j = J-1, \ldots, 0$, $k = 0, \ldots, 2^j - 1$ by the following rule:

set $\Delta_{j,k} = D(\{C_l(j,k,\cdot)\}_{l=1}^{L})$;

if $\Delta_{j,k} \geq \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$, that is, if discriminatory power of a parent node in WP tree is greater than those of children nodes;

then $A_{j,k} = B_{j,k}$;

else $A_{j,k} = A_{j+1,2k} \oplus A_{j+1,2k+1}$ and set $\Delta_{j,k} = \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$.

Step 5. Order the chosen base functions by their power of discrimination.

Step 6. Use k (normally much less than N) most discriminant base functions for constructing classifiers.

After Step 4 is performed, a complete set of orthogonal bases are constructed. Orthogonality of the bases ensures that wavelet coefficients used as features during classification process are uncorrelated as much as possible. Subsequently, one can use all the WP coefficients from each of the terminal nodes of the pruned tree or just use their subset with k highest discriminant bases in Step 6 or employ a statistical method to produce low-dimensional features as the input features of a classifier for discriminating different classes. In this paper, the WP coefficients of the selected optimal WP nodes are taken for calculating the root energy contained in each node. Mathematically, for the WP coefficients $W_{j,k}(i)$, $i = 1, \ldots, 2^{n_0-j}$, from each node (j,k), their root energy is calculated as

$$E_{j,k} = \sqrt{\sum_{i=1}^{2^{n_0-j}} W_{j,k}(i)^2}. \tag{6}$$

Totally, the root energy values of all of the selected nodes are put together to formulate a vector denoted by $\mathbf{E}_S$ (where $S = \{(j,k)\}$ is the subscript set of the selected WP nodes $W_{j,k}$) which is conveniently called the LDB feature. The set of root energy values contained in the WP nodes at the final level J of the WPT is called WPT feature in this paper and denoted by $\mathbf{E}_J$.

2.3. LPP for Feature Pattern Mining. In this section, we briefly describe the LPP algorithm of learning a locality preserving subspace from a high-dimensional data containing the sample values of a feature vector $\mathbf{E}_S$ or $\mathbf{E}_J$. Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{q\times n}$ denote a data matrix, representing a set of q-dimensional samples of size n with zero mean. Now, consider the problem of representing the data matrix $\mathbf{Y}$ by a single vector $\mathbf{x} = [x_1, \ldots, x_n]$ such that x_i represents $\mathbf{y}_i$. We will thus find a linear mapping, denoted by a transformation vector $\mathbf{w} \in \mathbb{R}^q$, from the q-dimensional space to a one-dimensional space, so that $\mathbf{w}^T\mathbf{y}_i = x_i$. LPP is a technique that seeks to preserve the intrinsic geometry of the data and local structure. The criterion of the objective function for choosing a map of the LPP is as follows:

$$\min \sum_{i,j} \left(x_i - x_j\right)^2 S_{ij}, \tag{7}$$

where x_i is the one-dimensional representation of $\mathbf{y}_i$ and the matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ is a similarity matrix.

A possible way of defining $\mathbf{S}$ is as follows:

$$S_{ij} = \begin{cases} \exp\left(\dfrac{-\left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2}{t}\right), & \left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2 < \varepsilon \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where parameter $t \in \mathbb{R}$ and ε defines the radius of the local neighborhood and is sufficiently small but bigger than 0. Two samples $\mathbf{y}_i$ and $\mathbf{y}_j$ are viewed within a local ε-neighborhood provided that $\|\mathbf{y}_i - \mathbf{y}_j\|^2 < \varepsilon$.

The objective function in (7) with the choice of symmetric weights S_{ij} ($S_{ij} = S_{ji}$) will be heavily penalized if neighboring points $\mathbf{y}_i$ and $\mathbf{y}_j$ are mapped far apart, that is, if $(x_i - x_j)^2$ is large. Therefore, minimizing the objective function is to ensure that if $\mathbf{y}_i$ and $\mathbf{y}_j$ are close, then x_i and x_j are close as well. Based on this point, the local structure of the input data can be preserved. Following some algebraic steps, we can get

$$\begin{aligned} &\frac{1}{2}\sum_{i,j}\left(x_i - x_j\right)^2 S_{ij} \\ &\quad = \frac{1}{2}\sum_{i,j}\left(\mathbf{w}^T\mathbf{y}_i - \mathbf{w}^T\mathbf{y}_j\right)^2 S_{ij} \\ &\quad = \sum_{i,j}\mathbf{w}^T\mathbf{y}_i S_{ij}\mathbf{y}_i^T\mathbf{w} - \sum_{i,j}\mathbf{w}^T\mathbf{y}_i S_{ij}\mathbf{y}_j^T\mathbf{w} \\ &\quad = \mathbf{w}^T\mathbf{Y}\left(\mathbf{D} - \mathbf{S}\right)\mathbf{Y}^T\mathbf{w} = \mathbf{w}^T\mathbf{Y}\mathbf{L}\mathbf{Y}^T\mathbf{w}, \end{aligned} \tag{9}$$

where $\mathbf{D}$ is a diagonal matrix; its entries are column (or row since $\mathbf{S}$ is symmetric) sums of $\mathbf{S}$, $D_{ii} = \Sigma_j S_{ij}$. $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the Laplacian matrix. The bigger the value D_{ii} (corresponding to x_i) is, the more important x_i is. Therefore, the LPP algorithm imposes a constraint as follows:

$$\mathbf{x}^T\mathbf{D}\mathbf{x} = 1 \Longrightarrow \mathbf{w}^T\mathbf{Y}\mathbf{D}\mathbf{Y}^T\mathbf{w} = 1. \tag{10}$$

Then, the minimization problem reduces to finding

$$\begin{aligned} &\underset{\mathbf{w}}{\operatorname{argmin}} \quad \mathbf{w}^T\mathbf{Y}\mathbf{L}\mathbf{Y}^T\mathbf{w} \\ &\text{s.t.} \qquad\quad \mathbf{w}^T\mathbf{Y}\mathbf{D}\mathbf{Y}^T\mathbf{w} = 1. \end{aligned} \tag{11}$$

The transformation vector $\mathbf{w}$ that minimizes the objective function is finally given by the minimum eigenvalue solution to the generalized eigenvalue problem:

$$\mathbf{w}^T\mathbf{Y}\mathbf{L}\mathbf{Y}^T\mathbf{w} = \lambda\mathbf{w}^T\mathbf{Y}\mathbf{D}\mathbf{Y}^T\mathbf{w}, \tag{12}$$

where the matrices $\mathbf{Y}\mathbf{L}\mathbf{Y}^T$ and $\mathbf{Y}\mathbf{D}\mathbf{Y}^T$ are symmetric and positive semidefinite. The top several projective vectors that minimize the objective function are the optimal linear approximations to the eigenfunctions of the Laplace Beltrami operator on the manifold, so they are capable of discovering the nonlinear manifold structure [33]. In this paper, the top several projective vectors are chosen as the mapping vectors to represent the LDB feature. They characterize the inherent class pattern and thus are hoped to mine the useful sensitive features for classification.

2.4. Proposed Feature Extraction Scheme for Data Classification. In the techniques mentioned above, the LDB and the LPP techniques have specific merits for classification. Specifically, the LDB algorithm focuses on identification of optimal decomposition subspaces for discriminatory feature extraction, while the LPP addresses the nonlinear pattern structure that represents the inherent condition class pattern. In other words, the LDB focuses on extraction of optimal raw features but each feature characterizes the class pattern in different sensitivity or local sensitivity, while LPP mainly addresses mining inherent class pattern feature embedded in the raw features. Therefore, this paper is proposed to combine the merits of these two techniques for a novel feature extraction. Specifically, the novel feature addresses extracting the inherent pattern structure embedded in the optimal WP nodes. Therefore, the proposed feature not only considers the static discriminatory WP node features themselves but also considers the dynamic sensitive class pattern structure embedded in the samples.

The idea of the proposed feature is illustrated in Figure 1. It can be found that although the optimal WP nodes (filled in black in Figure 1) have been selected through the LDB algorithm, they have different sensitivity in characterizing the class pattern. However, after conducting the LPP algorithm on the feature values, a new sensitive feature that clearly represents the class pattern is effectively extracted. In this process, the sensitive feature characterizes a nonlinear class pattern manifold embedded in sample values of the raw features. This indicates that LPP is beneficial to improve the class sensitivity of the selected

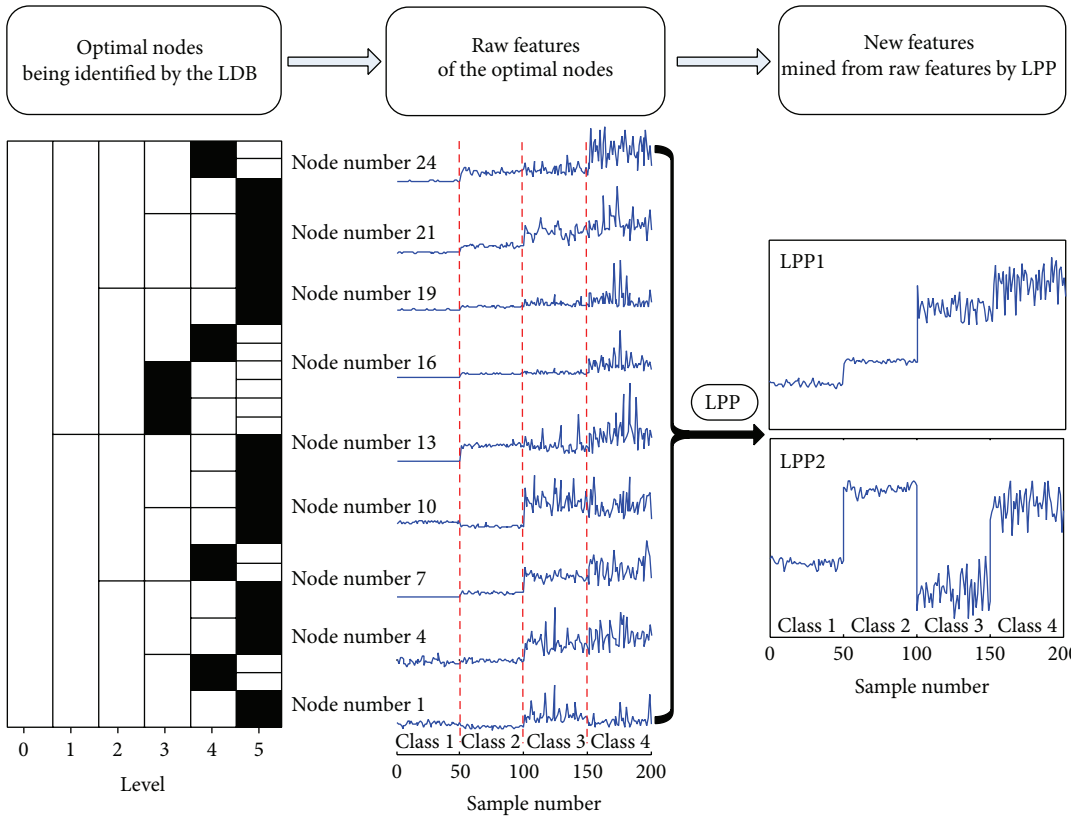


Figure 1: Principle of the proposed feature extraction scheme based on combination of LDB and LPP.

discriminatory features. Therefore, this new kind of feature can be well used to vibration data-based machine fault classification.

Based on the principle of combination of LDB and LPP, the proposed feature extraction algorithm can be then described as follows.

LDB-LPP Feature Extraction Algorithm. A training dataset consisting of L class of signals and a testing dataset are given.

Step 1. Conduct the WPT to decompose the signals contained in the dataset into the WP library with level J via (2).

If the signal is from the training dataset, go to Step 2.

Else, if the signal is from the testing dataset, then go to Step 3.

Step 2. Conduct the LDB algorithm to identify the optimal WP nodes that supply maximum dissimilarity information among the training dataset.

Step 3. Calculate the root energy of the coefficients of selected WP nodes to constitute a raw feature set $\mathbf{E}_S$ via (6).

If the signal is from the training dataset, go to Step 4.

Else, if the signal is from the testing dataset, then go to Step 5.

Step 4. Conduct the LPP algorithm to the raw feature value sets of the training dataset to obtain the mapping matrix through solving (12); then go to Step 5.

Step 5. Use the mapping matrix in Step 4 to calculate the new feature values of the dataset.

The proposed features have the most sensitive discriminatory capability and are thus chosen as inputs to a diagnostic classifier for characterizing data classes. To make it clearer, the flowchart of the proposed algorithm is shown in Figure 2 as well as the scheme of machine fault classification. The machine fault classification scheme includes two parts: the LDB-LPP feature values are firstly extracted for both the training and testing signals and then a diagnostic classifier is trained for classification of the fault signals.

3. Experimental Results and Analysis

In order to evaluate the effectiveness of the feature extraction scheme proposed above for machine fault classification, the bearing data with multiple faults from real bearing experiments are analyzed in this study.

3.1. Experimental Dataset. The experimental data are from Case Western Reserve University Bearing Data Center [34]. The experimental setup consists of four parts which are

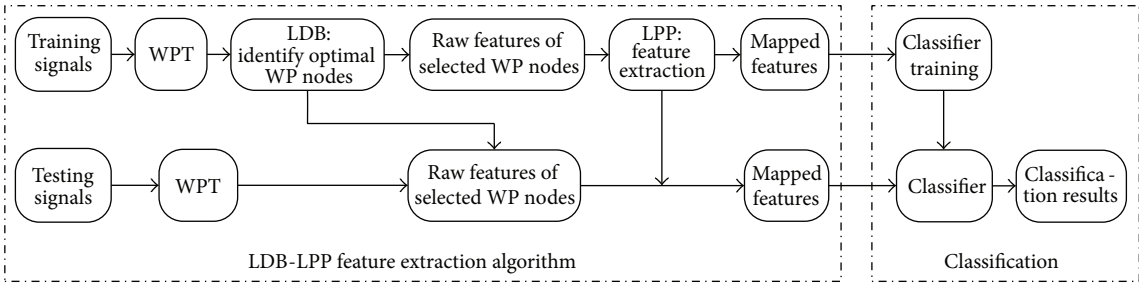


Figure 2: Flowchart of the proposed LDB-LPP feature extraction algorithm and machine fault classification.

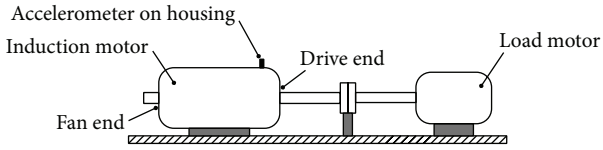


Figure 3: A schematic of the experimental system.

Table 1: Description of datasets for bearing data classification.

Dataset	Number of samples (training/testing)	Fault type	Fault size (inch)	Label of class
A	50/50	Normal	0	1
		Ball	0.007	2
		Inner race	0.007	3
		Outer race	0.007	4
		Ball	0.014	5
		Inner race	0.014	6
		Outer race	0.014	7
B	25/25	Ball	0.021	8
		Inner race	0.021	9
		Outer race	0.021	10

an induction motor, a dynamometer, a torque transducer, and control electronics. The resulting vibration was measured by an accelerometer being mounted to the motor housing at the drive end of the motor as illustrated in Figure 3. The accelerometer is a vibration sensor with a bandwidth up to 5000 Hz and a 1 V/g output. Single point faults of size 0.007, 0.014, 0.021, and 0.028 inches were set on the drive-end bearings by using the electric discharge machining approach. These faults were set, respectively, on rolling element, inner raceway, and outer raceway in the experiments. The sampling frequency of the data is 12 kHz with the sample length being 2000 and the motor speed was 1748 rev/min.

Datasets A and B to be analyzed consist of ten classes (class labels are marked in Table 1) covering different bearing fault types and severities as listed in Table 1. In the datasets, there are four different fault types including normal, outer-race fault, inner-race fault, and ball fault, and each of the last three fault types includes three different defect sizes of 0.007, 0.014, and 0.021 inches, respectively. In dataset A, the samples are split into 500 training ones (50 in each class) and 500 testing ones (50 in each class), while dataset B contains 250 training samples (25 in each class) and 250 testing samples (25 in each class). This is a complex ten-class problem to identify both the fault type and the fault size for the operating bearing conditions.

3.2. Feature Evaluation. In this study, the decomposition level of the WPT is set to be 6 and the Daubechies 8 wavelet is employed. The selected nodes by the LDB are shown in Figure 4. In the following study, the root energy of a signal decomposed into each selected node is calculated as the raw features in the proposed study, while that in each node at the last layer is used to form the traditional WPT feature for a comparison.

To quantitatively evaluate the capability of LDB feature in pattern classification, three common clustering evaluation metrics are analyzed as follows. The first is a widely used discriminant factor. Suppose that there is a feature vector $\{f_1, f_2, \ldots, f_d\}$, where d is the dimension of feature; then the discriminant factor is defined as follows:

$$S = \frac{S_B}{S_W}, \tag{13}$$

where S_B indicates the between-class scatter to describe the scattered level among different classes, while S_W is the within-class scatter which represents the concentrated level in the same classes. These two scatters are, respectively, defined as

$$S_B = \sum_{l=1}^{L} N_l \left\| \mu_f^l - \overline{\mu}_f \right\|,$$
$$S_W = \sum_{l=1}^{L} \sum_{j=1}^{N_l} \left\| f_j - \mu_f^l \right\|, \tag{14}$$

where N_l is the total number and μ_f^l is the average feature vector for samples in the lth class and $\overline{\mu}_f$ is the total average of the feature vectors for all classes. It can be seen that the discriminant factor S is a comprehensive indicator that combines between-class scatter and within-class scatter. A larger discriminant factor is better for classification purpose to characterize the discriminating capability of the given feature.

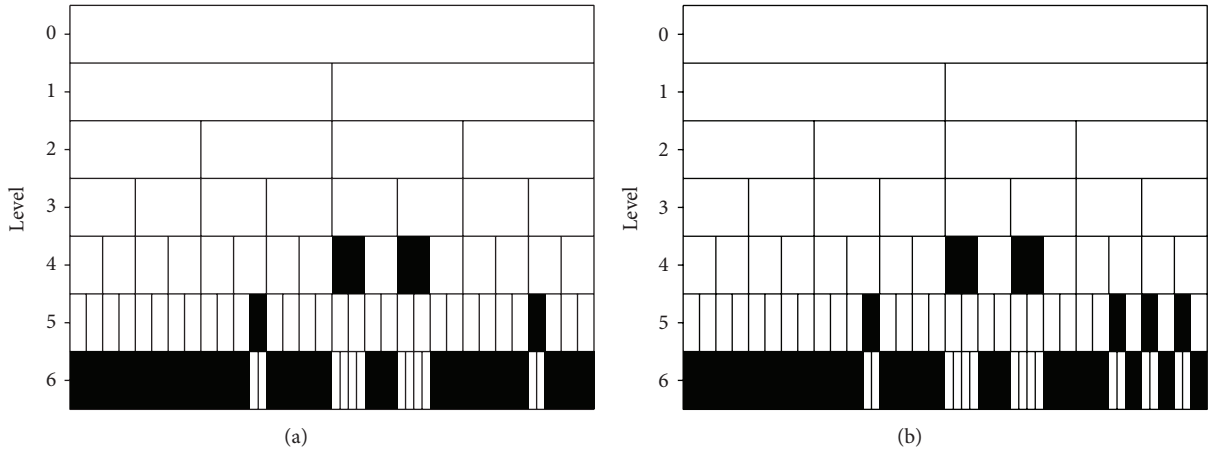


Figure 4: Optimal WP nodes selected by the LDB at level 6: (a) dataset A and (b) dataset B.

The other two clustering evaluation metrics are the cluster accuracy (ACC) and normalized mutual information (NMI) metrics [35], which are defined as follows, respectively.

Cluster Accuracy (ACC). Assuming that s_i and r_i are the acquired new label and the provided real label of the given point μ_i, the ACC is defined as follows:

$$\mathrm{ACC} = \frac{\sum_{i=1}^{Ns} \delta\left(r_i, \mathrm{Map}\left(s_i\right)\right)}{Ns}, \qquad \delta\left(p,q\right) = \begin{cases} 1 & p = q \\ 0 & \text{otherwise}, \end{cases} \tag{15}$$

where Ns is the total number of the samples and $\mathrm{Map}(s_i)$ is the optimal mapping function that ranges each s_i to match the real label and can be found by the Kuhn-Munkres (KM) algorithm. Here, we assume that the relationship of the identified clusters with the predefined classes is known. Thus, it is easy to imagine that a larger ACC value indicates better clustering and generally better classification.

Normalized Mutual Information (NMI). It is a mutual information (MI) metric and defined as

$$\mathrm{NMI} = \frac{\sum_{i=1}^{L}\sum_{j=1}^{L} t_{i,j} \log\left(\left(Ns \cdot t_{i,j}\right)/t_i \tilde{t}_j\right)}{\sqrt{\sum_{i=1}^{L} t_i \log\left(t_i/Ns\right)}\sqrt{\sum_{j=1}^{L} \tilde{t}_j \log\left(\tilde{t}_j/Ns\right)}}, \tag{16}$$

where t_i is the number of the acquired samples in class C_i and $\tilde{t}_j$ is the number of the provided samples in the ground truth class C_j. In addition, $t_{i,j}$ is the number of the intersected samples between class C_i and class C_j. A larger NMI reveals better clustering performance, which is beneficial to classification.

For a visible purpose and a fair comparison, the dimensions of the LDB and the WPT features are both reduced to 3 by using dimensionality reduction techniques including the LPP and the traditional PCA. Note that the LDB feature followed by the LPP just generates the proposed feature in this study. We then calculated the mentioned three clustering evaluation metrics. Here, k-means clustering method is applied to obtain the cluster label in the reduced 3-dimentional features before calculating the ACC and NMI. The k number used in the k-means clustering method is set as L which is the number of the class. What is more, to realize efficient and stable convergence, we set the L initial points as the intermediate point of each class in mathematics.

Table 2: Clustering evaluation of different features for dataset A.

Metric	PCA		LPP	
	WPT	LDB	WPT	LDB
S	3.0900	4.0740	7.8135	8.7864
ACC	0.4700	0.7560	0.9740	0.9800
NMI	0.7914	0.9090	0.9546	0.9623

The neighborhood parameter of LPP is taken as 12. As an illustration, the scatter plots of the ten-class dataset A for training data are drawn in Figure 5. It can be seen that the LDB feature shows a better classification capability than the WPT features. On the other hand, it can be also found that the LPP has a much more excellent classification capability than the PCA. Therefore, the LDB-LPP shows the best classification capability in the between-class and within-class scatter performance. The extracted feature patterns are also demonstrated in Figure 6, where it can be clearly seen that the third LPP of LDB feature values characterizes a better difference for each class as compared to the third LPP of WPT feature values. Moreover, the quantitative results as listed in Table 2 also support the above statements. It can be seen that the clustering evaluation metrics S, ACC, and NMI of the LDB feature are higher than those of the WPT feature, and LPP performs much better than PCA. The combination of LDB and LPP shows the most beneficial performance for classification. Moreover, the clustering evaluation of dataset B (with half the number of samples of dataset A) is also computed here as shown in Table 3. These clustering evaluation values show the same tendency and indicate that the LPP can learn a good nonlinear class pattern structure among the discriminatory LDB feature values.

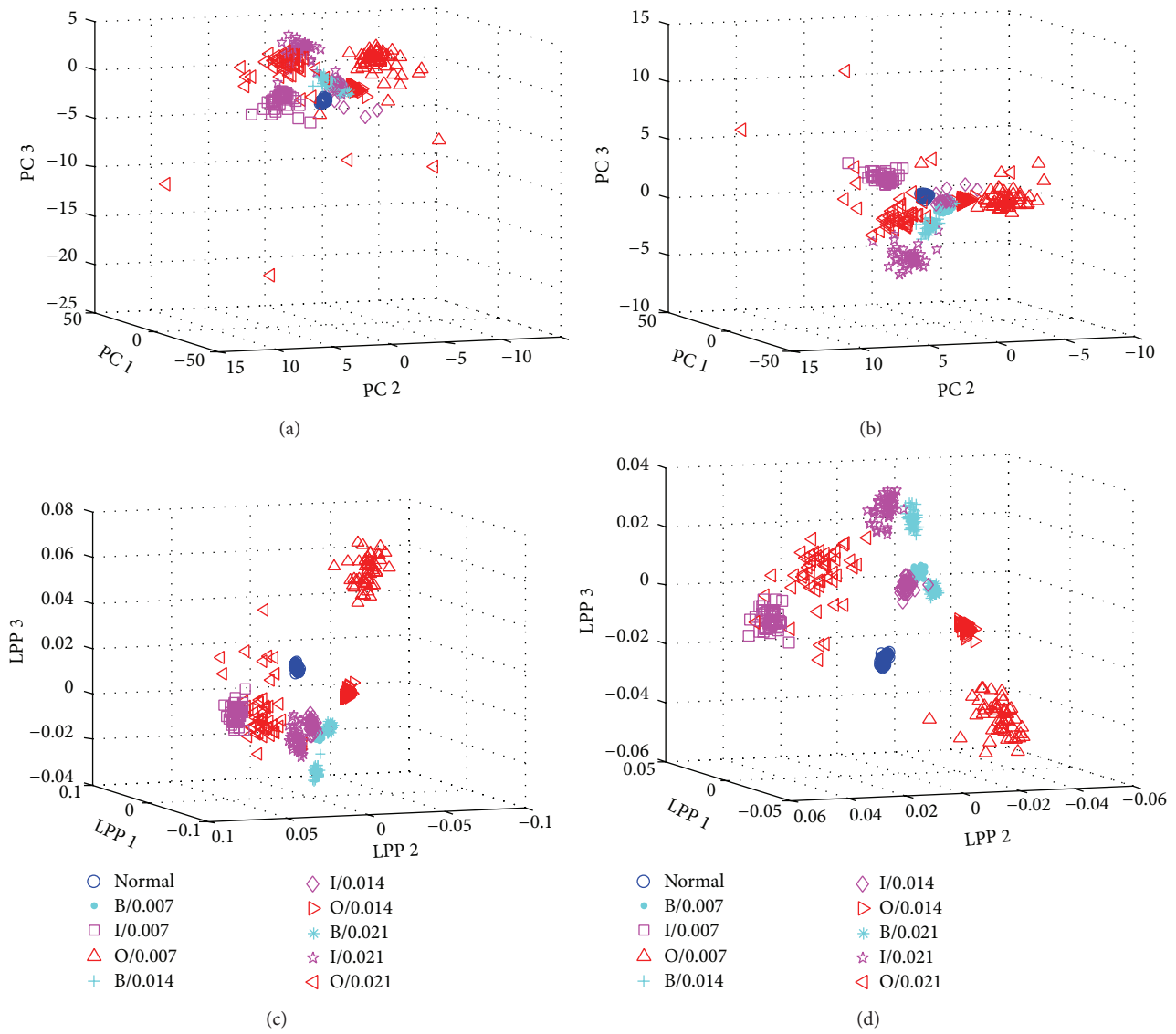


Figure 5: Representation of training samples of bearing dataset A: (a) PCA of WPT feature values, (b) PCA of LDB feature values, (c) LPP of WPT feature values, and (d) LPP of LDB feature values.

Table 3: Clustering evaluation of different features for dataset B.

Metric	PCA		LPP	
	WPT	LDB	WPT	LDB
S	3.0260	3.7721	9.5638	9.7630
ACC	0.5120	0.7720	0.9280	1
NMI	0.779	0.8149	0.9040	1

3.3. Classification of Fault Types and Severities. To further evaluate the performance of the proposed feature in data classification, the ten-class datasets A and B are employed for fault classification by comparing various features. In this study, the proposed LDB-LPP feature is compared to traditional feature extraction methods including PCA, LDA, LPP, supervised LPP (SLPP) [36], LE, and LLE. Among the six methods, PCA and LPP are unsupervised linear techniques, LDA and SLPP are supervised linear techniques, and LE and LLE are nonlinear manifold learning techniques.

To emphasize the feature performance, the nearest mean classifier, one of the simplest and the most intuitive statistical classifiers, is applied for classification in this study. This classifier is based on the principle of the closest Euclidean distance and the concept of similarity that similar patterns should be assigned to the same class. In this study, the mean vector of the training data in each class is used to represent each pattern class. Patterns of samples can be distinguished according to the minimum distance criterion which means maximum similarity. Moreover, another advanced classifier, Gaussian mixture model (GMM) classifier, is also applied in this study.

The recognition accuracy of the proposed LDB-LPP feature and the other comparison features (extracted from the WPT feature without node selection) are shown in Table 4. It can be seen that the proposed LDB-LPP feature

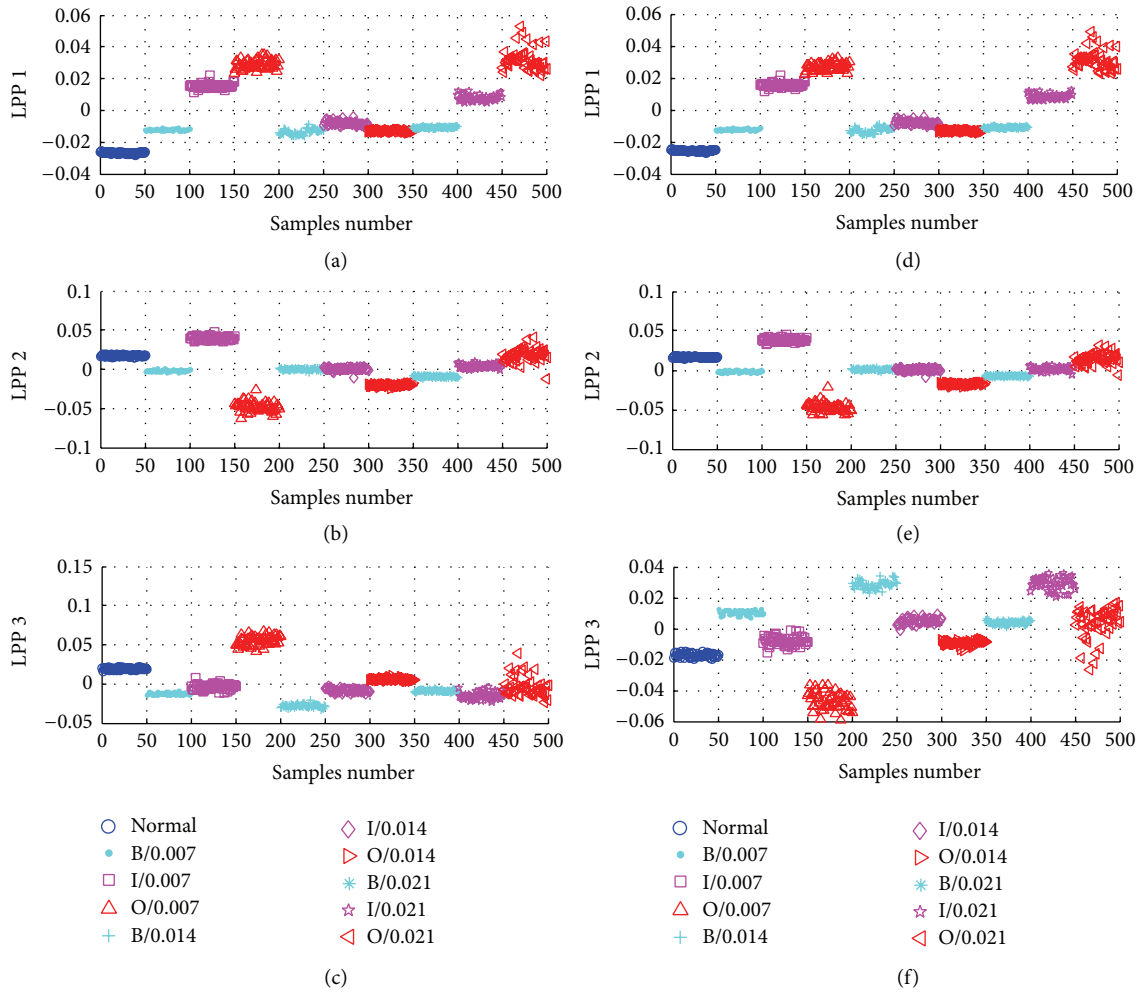


Figure 6: The first three projections of training samples of bearing dataset A: (a~c) LPPs of WPT feature values and (d~f) LPPs of LDB feature values.

Table 4: Recognition accuracy (%) performance for dataset A and dataset B.

Classifier	Feature Extraction	Dataset A		Dataset B	
		Training	Testing	Training	Testing
Nearest Mean	PCA	84.6	84.2	85.6	82
	LPP	98.8	99	98.8	96
	SLPP	98.6	99	99.6	97.6
	LE	97.8	97.2	95.2	90.4
	LLE	97.2	98.2	98	98.4
	LDA	98.6	98.2	100	98.4
	LDB-LPP	99.8	100	99.2	99.6
GMM	PCA	97.8	90	96.8	83.6
	LPP	99.2	99	99.6	97.6
	SLPP	99.2	98.8	100	97.6
	LE	89	88.4	96.4	90.8
	LLE	98.8	99.2	100	99.6
	LDA	99.2	97.8	100	96.8
	LDB-LPP	100	100	100	100

outperforms the traditional features achieved by PCA, LDA, LPP, SLPP, LE, and LLE, which verifies the benefits of the LDB for choosing discriminatory features. Moreover, the GMM classifier further improves the recognition rate of the nearest mean classifier. Note that the recognition accuracy of dataset A is generally higher than that of dataset B because the number of samples in dataset A is bigger. It can be found that the promotion of the recognition accuracy of the LDB-LPP feature in comparison with the other features becomes more obvious in dataset B than in dataset A for two classifiers. For instance, dataset A shows an average promotion 1% for testing by considering the LDB in the LPP feature extraction, while dataset B displays an average promotion 3% (for two classifiers) for testing. Figures 7 and 8 intuitively display that the proposed LDB-LPP feature performs the best among all the comparison features. In this study, the testing recognition accuracy based on LDB-LPP feature is equal to or very close to 100% for two classifiers. These results imply that the proposed LDB joint LPP feature extraction method could obtain significant achievements in improving classification accuracy.

4. Conclusions

This paper presents a feature extraction method which integrates the LDB and the LPP to explore the useful and

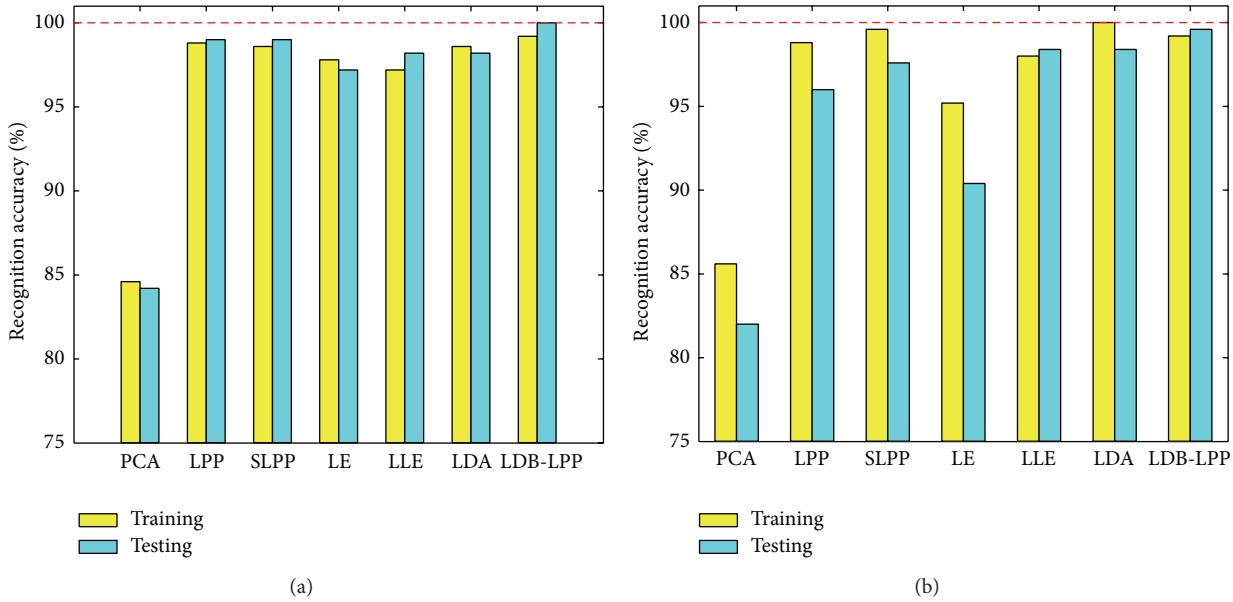


Figure 7: Recognition accuracy of ten-class classification by the nearest mean classifier: (a) dataset A and (b) dataset B.

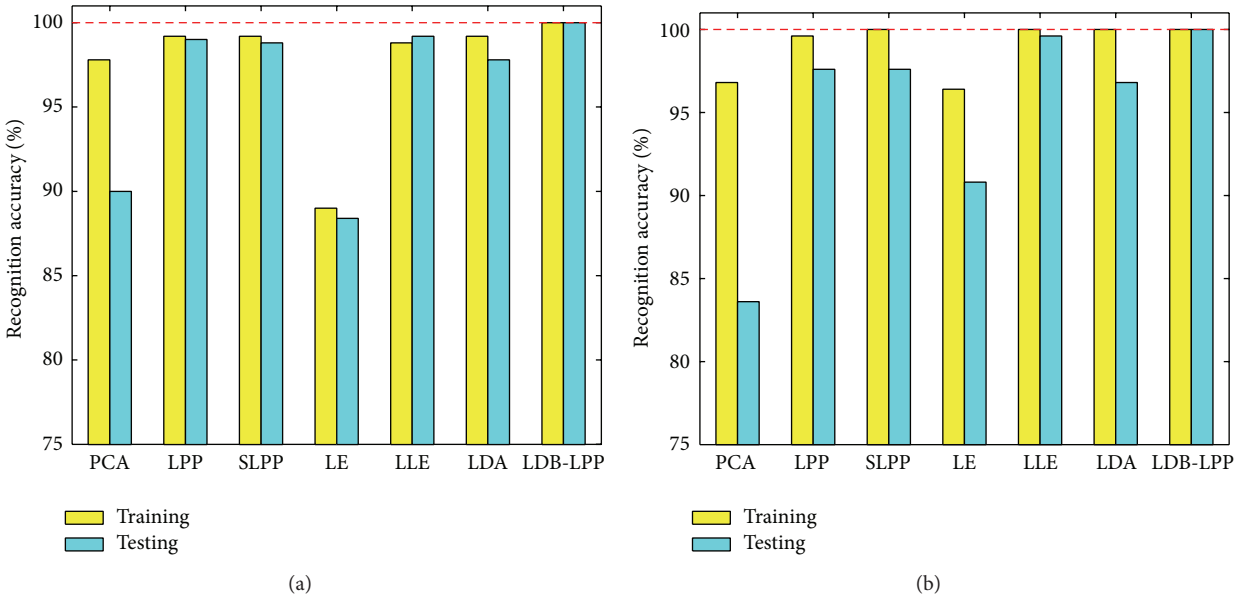


Figure 8: Recognition accuracy of ten-class classification by GMM classifier: (a) dataset A and (b) dataset B.

powerful characteristics for vibration data-based machine fault classification. The LDB is used to select the most discriminant WP bases from a library of redundant and orthogonal time-frequency subspaces. The input features are produced by the selected optimal wavelet bases but they possess different sensitivity in characterizing class information. The LPP is then employed to acquire the sensitive feature that characterizes the inherent class pattern feature embedded in the raw features for a much better identification accuracy. The proposed feature extraction method combines the merits of the LDB and the LPP and thus displays valuable benefits for data classification. To verify the effectiveness of the proposed method, the vibration data representing different bearing fault types and severities are analyzed by comparing with other features extracted from the WPT feature. The experimental results for bearing fault classification indicate that the LDB-LPP feature is more effective than those feature extraction methods based on the WPT feature without base selection. The presented LDB joint LPP feature extraction method is also hoped to be well-suited to other machine fault classification, such as gears, spindles, and cutting tools, due to the excellent feature representation for the class patterns.

Moreover, the technical aspects in the proposed LDB-LPP feature extraction framework can be further improved and

strengthened. First, this paper fairly compares the LDB feature and WPT feature in the same decomposition level, which can validate the benefits of the LDB in data classification. However, how to select the well-suited decomposition level in the LDB is still an open issue in the further study. Second, LPP is a typical and effective feature extraction method which obtains the manifold structure in a linear projection. Although the LPP has been successfully used to overcome the weakness of the LDB in this study, it is meaningful to apply the new well-performed manifold learning methods instead of LPP in the proposed framework to further enhance the performance of data classification. This should also depend on how complex the data to be analyzed is. Other possible applications on complex classification remained to be studied in the future.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 51005221), the Research Fund for the Doctoral Program of Higher Education of China (Grant no. 20103402120017) and the Program for New Century Excellent Talents in University, China (Grant no. NCET-13-0539). The authors would like to thank Case Western Reserve University for offering free download of the bearing data and the anonymous reviewers for their constructive and valuable comments.

References

[1] H. Qiu, J. Lee, J. Lin, and G. Yu, "Robust performance degradation assessment methods for enhanced rolling element bearing prognostics," *Advanced Engineering Informatics*, vol. 17, no. 3-4, pp. 127–140, 2003.

[2] Q. He, F. Kong, and R. Yan, "Subspace-based gearbox condition monitoring by kernel principal component analysis," *Mechanical Systems and Signal Processing*, vol. 21, no. 4, pp. 1755–1772, 2007.

[3] J. Yu, "Bearing performance degradation assessment using locality preserving projections," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7440–7450, 2011.

[4] Y. Lei, J. Lin, Z. He, and D. Kong, "A method based on multisensor data fusion for fault detection of planetary gearboxes," *Sensors*, vol. 12, no. 2, pp. 2005–2017, 2012.

[5] Z. Liu, X. Chen, Z. He, and Z. Shen, "LMD method and multiclass RWSVM of fault diagnosis for rotating machinery using condition monitoring information," *Sensors*, vol. 13, no. 7, pp. 8679–8694, 2013.

[6] S. Wang, X. Sun, and C. Li, "Wind turbine gearbox fault diagnosis method based on Riemannian manifold," *Mathematical Problems in Engineering*, vol. 2014, Article ID 153656, 10 pages, 2014.

[7] R. Yan, R. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: a review with applications," *Signal Processing*, vol. 96, pp. 1–15, 2014.

[8] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713–718, 1992.

[9] N. Saito and R. R. Coifman, "Local discriminant bases and their applications," *Journal of Mathematical Imaging and Vision*, vol. 5, no. 4, pp. 337–358, 1995.

[10] E. Hulata, R. Segev, Y. Shapira, M. Benveniste, and E. Ben-Jacob, "Detection and sorting of neural spikes using wavelet packets," *Physical Review Letters*, vol. 85, no. 21, pp. 4637–4640, 2000.

[11] N. Saito, R. R. Coifman, F. B. Geshwind, and F. Warner, "Discriminant feature extraction using empirical probability density estimation and a local basis library," *Pattern Recognition*, vol. 35, no. 12, pp. 2841–2852, 2002.

[12] D. J. Strauss, G. Steidl, and W. Delb, "Feature extraction by shape-adapted local discriminant bases," *Signal Processing*, vol. 83, no. 2, pp. 359–376, 2003.

[13] D. Li, W. Pedrycz, and N. J. Pizzi, "Fuzzy wavelet packet based feature extraction method and its application to biomedical signal classification," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 6, pp. 1132–1139, 2005.

[14] R. Tafreshi, F. Sassani, H. Ahmadi, and G. Dumont, "Local discriminant bases in machine fault diagnosis using vibration signals," *Integrated Computer-Aided Engineering*, vol. 12, no. 2, pp. 147–158, 2005.

[15] K. Umapathy and S. Krishnan, "Modified local discriminant bases algorithm and its application in analysis of human knee joint vibration signals," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 517–523, 2006.

[16] K. Umapathy, S. Krishnan, and R. K. Rao, "Audio signal feature extraction and classification using local discriminant bases," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1236–1246, 2007.

[17] Q. He, R. Yan, and R. X. Gao, "Wavelet packet base selection for gearbox defect severity classification," in *Proceedings of the Prognostics and System Health Management Conference (PHM '10)*, Macau, China, January 2010.

[18] D. Kim, J. J. Liu, and C. H. Han, "Determination of steel quality based on discriminating textural feature selection," *Chemical Engineering Science*, vol. 66, no. 23, pp. 6264–6271, 2011.

[19] Q. He, R. Yan, F. Kong, and R. Du, "Machine condition monitoring using principal component representations," *Mechanical Systems and Signal Processing*, vol. 23, no. 2, pp. 446–466, 2009.

[20] R. Yan and R. Gao, "Wavelet domain principal feature analysis for spindle health diagnosis," *Structural Health Monitoring*, vol. 10, no. 6, pp. 631–642, 2011.

[21] L. H. Chiang, E. L. Russell, and R. D. Braatz, "Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 2, pp. 243–252, 2000.

[22] X. Jin, M. Zhao, T. Chow, and M. Pecht, "Motor bearing fault diagnosis using trace ratio linear discriminant analysis," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 5, pp. 2441–2451, 2014.

[23] A. Widodo and B. Yang, "Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors," *Expert Systems with Applications*, vol. 33, no. 1, pp. 241–250, 2007.

[24] C. Sun, Z. Zhang, Z. He, Z. Shen, B. Chen, and W. Xiao, "Novel method for bearing performance degradation assessment? A

kernel locality preserving projection based approach," *Proceedings of the Institution of Mechanical Engineers C: Journal of Mechanical Engineering Science*, vol. 228, pp. 548–560, 2013.

[25] Q. Jiang, M. Jia, J. Hu, and F. Xu, "Machinery fault diagnosis using supervised manifold learning," *Mechanical Systems and Signal Processing*, vol. 23, no. 7, pp. 2301–2311, 2009.

[26] Q. He, "Time-frequency manifold for nonlinear feature extraction in machinery fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 35, no. 1-2, pp. 200–218, 2013.

[27] Z. Su, B. Tang, J. Ma, and L. Deng, "Fault diagnosis method based on incremental enhanced supervised locally linear embedding and adaptive nearest neighbor classifier," *Measurement*, vol. 48, pp. 136–148, 2014.

[28] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[29] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[30] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[31] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM Journal on Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2005.

[32] M. H. C. Law and A. K. Jain, "Incremental nonlinear dimensionality reduction by manifold learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 377–391, 2006.

[33] X. He and P. Niyogi, "Locality preserving projections," in *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS '03)*, vol. 16, pp. 234–241, Vancouver, Canada, December 2003.

[34] Bearing data center, http://csegroups.case.edu/bearingdatacenter/home.

[35] M. Zhao, X. Jin, Z. Zhang, and B. Li, "Fault diagnosis of rolling element bearings via discriminative subspace learning: visualization and classification," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3391–3401, 2014.

[36] D. Cai, X. He, and J. Han, "Using graph model for face analysis," Tech. Rep. no. 2636, 2005.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 819438, 8 pages
http://dx.doi.org/10.1155/2014/819438

Research Article

A New Feature Selection Algorithm Based on the Mean Impact Variance

Weidong Cheng,[1] Tianyang Wang,[1] Weigang Wen,[1] Jianyong Li,[1] and Robert X. Gao[2]

[1] *School of Mechanical Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China*
[2] *Department of Mechanical Engineering, University of Connecticut, Storrs, CT 06269, USA*

Correspondence should be addressed to Weidong Cheng; wdcheng@bjtu.edu.cn

Received 19 January 2014; Revised 1 May 2014; Accepted 9 June 2014; Published 26 June 2014

Academic Editor: Weihua Li

The selection of fewer or more representative features from multidimensional features is important when the artificial neural network (ANN) algorithm is used as a classifier. In this paper, a new feature selection method called the mean impact variance (MIVAR) method is proposed to determine the feature that is more suitable for classification. Moreover, this method is constructed on the basis of the training process of the ANN algorithm. To verify the effectiveness of the proposed method, the MIVAR value is used to rank the multidimensional features of the bearing fault diagnosis. In detail, (1) 70-dimensional all waveform features are extracted from a rolling bearing vibration signal with four different operating states, (2) the corresponding MIVAR values of all 70-dimensional features are calculated to rank all features, (3) 14 groups of 10-dimensional features are separately generated according to the ranking results and the principal component analysis (PCA) algorithm and a back propagation (BP) network is constructed, and (4) the validity of the ranking result is proven by training this BP network with these seven groups of 10-dimensional features and by comparing the corresponding recognition rates. The results prove that the features with larger MIVAR value can lead to higher recognition rates.

1. Introduction

Feature extraction is key factor in pattern recognition because only sufficient and effective features can describe a given sample comprehensively and then differentiate between classes [1–4]. In general, there always exist tens or hundreds of variables to describe an object. However, the use of too many features in pattern recognition is not suitable for the following reasons: first, the number of feature dimensions should be far fewer than the number of training sets; second, too many features increase training and utilization times, which then cause the entire recognition algorithm to be time-consuming; last, prediction performance is negatively affected by inappropriate features. Because of these problems, the algorithms for multivariate feature selection and feature ranking have become the focus of much research in several areas [5].

Feature selection is a necessary preprocessing step between feature extraction and pattern recognition. Its main purpose is to choose more sensitive features from the original multidimensional features as the subset that should maintain the same ability of recognition. To achieve this goal, several algorithms based on the principal component analysis (PCA), artificial neural network (ANN), genetic algorithm (GA), support vector machine (SVM), and pattern recognition theory-based algorithm are proposed. The PCA algorithm is the most common linear dimensionality reduction algorithm that can map multidimensional features into a space of lower dimension. Reference [6] employs this algorithm in face recognition and [7] makes use of it in machine defect classification. However, PCA can only lower the dimension by generating new features that are not suitable if the physical meaning of the features must be given [8]. As intelligent algorithms, the ANN and GA algorithms can be used in feature selection. Among them, an ANN-based feature-selection method called the UTA algorithm (named after the author [9]) is used to predict the American business cycle. The GA algorithm is used to select features for SVM [10].

However, the GA algorithm is too complicated and cannot quantitatively determine the feature that is more suitable for classification [11]. Reference [12] proposed a recursive SVM feature selection for mass-spectrometry and microarray data. Another pattern recognition theory-based algorithm was proposed in [13]. Its main principle is to maximize the quotient obtained by dividing the mean distance between the samples of different classes by the mean distance between the samples of the same classes. This method is widely used in parameter evaluation [14] because of its efficiency and clear mathematic meaning.

In this study, an interesting method called the mean impact variance (MIVAR) method is constructed to determine the feature that is more sensitive to classification. This method is obtained after the BP network training step by changing the magnitude of all the features separately. The feature with the larger MIVAR value is considered the better choice when the BP network is used as the classifier. To verify the effectiveness of this method, we use it to rank multidimensional time-domain features and select more representative features for a bearing fault diagnosis.

The rest of the paper is organized as follows: Section 2 specifies the algorithm of the MIVAR-based feature selection; Section 3 describes the databases and the all-waveform feature extraction method, which is used to generate multidimensional features; Section 4 uses the MIVAR method to rank the aforementioned multidimensional features in the order of their sensitivity and the BP network to testify the validity of the rank result. Finally, the conclusion is presented in Section 5.

2. MIVAR-Based Feature Selection Algorithm

MIVAR is a new method that can be used to select more representative features from multidimensional features. To specify the algorithm in detail, n is used to represent the number of classes, L is used to represent the total sample number of the in the training set, and l is used to represent the number of each class ($L/l = n$). The dimension of the multidimensional feature is m. i, j, and k represent the feature sequence number, the class, and the sample in one class, respectively. The specific algorithm is described as follows.

Step 1. First, m-dimensional features are extracted from the training sets of n different classes. A BP network is then constructed and trained with the training sets. The input size of the network is m, which is equal to the dimension of the multidimensional features. The output size is equal to n, which represents the type number.

Step 2. The kth sample is chosen from the training sets of the jth class, and the results are obtained by feeding the trained BP network with the corresponding m-dimensional feature. Then, the value of the ith dimension varied by ± 30% (the other $m-1$ dimensions are maintained at the same values) to form the following two new features:

$$P_{i,k,j,\mathrm{UP}} = P_{i,k,j} \cdot (1 + 30\%), \quad (1)$$

$$P_{i,k,j,\mathrm{DOWN}} = P_{i,k,j} \cdot (1 - 30\%), \quad (2)$$

where i is the dimension sequence number, k is the sequence number of the sample in the jth class, UP means that the new feature is generated by increasing the value of the ith dimension by 30%, and DOWN means that the new feature is generated by decreasing the value of the ith feature by 30%. $P_{i,k,j}$ is the original feature and $P_{i,k,j,\mathrm{UP}}$ and $P_{i,k,j,\mathrm{DOWN}}$ are the two new generated features. Except for the ith feature, the other $n-i$ pairs of new features should be calculated also following (1) and (2).

Step 3. The network is simulated with these n pairs of new features, and n pairs of outputs, $O_{i,k,j,\mathrm{UP}}$ and $O_{i,k,j,\mathrm{DOWN}}$, where i varies from one to n, are obtained.

Step 4. The absolute value of the difference between the jth bits of $O_{i,k,j,\mathrm{UP}}$ and $O_{i,k,j,\mathrm{DOWN}}$ is calculated. Here, we use $\mathrm{IV}_{i,k,j}$ to denote the difference, which represents how much the ith feature affects the correct recognition of the kth sample of the jth class. $O_{i,k,j,\mathrm{UP}}$ and $O_{i,k,j,\mathrm{DOWN}}$ are both $j \times 1$ matrices, and the jth bit can determine whether the kth sample belongs to the jth class. We call the jth bit the judging bit of the jth class. Consider

$$\mathrm{IV}_{i,k,\mathrm{NO}} = \left| O_{i,k,j,\mathrm{UP}} - O_{i,k,j,\mathrm{DOWN}} \right|. \quad (3)$$

Step 5. The process is repeated from Step 2 to Step 4 for the other $l-1$ samples of the jth class, and another set of $l-1$ differences of the ith feature is obtained for the jth class. In addition to the $\mathrm{IV}_{i,k,\mathrm{NO}}$ obtained in Step 4, we have a total of l differences of the ith feature. By calculating the mean value of these m differences, we obtain MIV, which represents how much the ith feature influences the correct recognition of a sample of the ith as follows:

$$\mathrm{MIV}_{i,\mathrm{NO}} = \frac{1}{l} \sum_{k=1}^{l} \mathrm{IV}_{i,k,\mathrm{NO}}. \quad (4)$$

Step 6. The process is repeated from Step 2 to Step 5 using the samples that belong to $n-1$ classes. This way, we can obtain all the MIVs of every feature for the samples of four different states: $\mathrm{MIV}_{i,1}, \ldots, \mathrm{MIV}_{i,j}, \ldots, \mathrm{MIV}_{i,n}$.

Step 7. The variance of the four MIVs of each feature is calculated for the four different states, and a method called MIVAR, which represents the fluctuation in the MIVs, is obtained. Consider

$$\mathrm{MIVAR}_i = \frac{\sum_{j=1}^{n} \left(\mathrm{MIV}_{i,j} - \overline{\mathrm{MIV}_i} \right)^2}{n}. \quad (5)$$

MIVAR is a proposed method that can determine the feature that is more suitable for classification. Thus, we should select a feature with a larger MIVAR as the one for final classification.

3. Database Description and Features Generation

In this paper, the effectiveness of the MIVAR-based feature selection algorithm is proven by selecting more representative

Table 1: Top three MIV sequence numbers for different classes.

Ordinal number	Top NO sequence numbers	Top IR sequence numbers	Top RE sequence numbers	Top OR sequence numbers
1	33	1	1	1
2	32	2	3	2
3	9	10	6	3

features for a bearing fault diagnosis using the data from the Bearing Data Center of Case Western Reserve University [15]. In detail, there are four data classes: normal (NO), inner race fault (IR), rolling element fault (RE), and outer race fault (OR). We choose 300 samples as the network training sets, and the sample number of each state is 75. In similar way, we choose another 100 samples as the testing sets that consist of 25 samples for NO, 25 samples for IR, 25 samples for RE, and 25 samples for OR. To acquire multidimensional features, a new feature-extraction method called all waveform feature extraction is proposed.

Step 1. The raw signal is rounded to the nearest hundredth, and the original signal data are divided into N groups (ensuring that the data from the same group are equal to each other). Then, each group number is counted and denoted with x_i, where i ranges from one to N.

Step 2. P_i represents the proportion of the ith group data number to the original signal total number, and it is obtained as follows:

$$P_i = \frac{x_i}{N}. \tag{6}$$

Step 3. The P_i curve is the probability density curve of the original signal. The four curves in Figure 1 represent the probability density curves of the vibration signal of NO, an IR, a RE, and an OR, respectively.

Step 4. New features are extracted on the basis of the probability density curve. The corresponding y-axis represents the percentage of each number in the different groups; thus, its upper bound is 100%. We choose 1/1,000 as the unit, equally divide the entire y-axis into 1,000 parts, and draw 1,000 lines parallel to the x-axis from every point along the y-axis. These 1,000 secants can be divided into two types, which are illustrated in Figure 2: the first type of secant intersects the curve more than two times (indicated by the lower two solid lines), and its corresponding features are equal to the distance between the intersections on the far right and the far left. The upper dotted line represents the second type of secant that has one or no intersections with the curve, and we let the feature obtained by this secant type be equal to zero. We let d_i denote the feature generated by the ith secant line $i = (1,2,\ldots,1{,}000)$, which is the all waveform feature.

While extracting the all waveform features of the training and testing sets, we find that the maximum value of P_is for all features is always less than 7% in all the training sets and testing sets. Therefore, we decrease the dimension of the all waveform feature from 1,000 to 70. Figure 3 shows how to obtain 70-dimensional features using a sample of NO: the lower two lines are the first secant and the 30th secant that belong to the first type. The bold solid lines in the middle represent the all waveform features, number 1 feature and number 30 feature, extracted by these two secant lines. The upper line that has no interaction with the curve belongs to the second secant type. It can generate number 70 feature whose value is equal to zero. Using the method described above, we extracted the 70-dimensional feature vector as the feature that represents the bearing vibrational signal of the training sets and testing sets.

4. Effectiveness Proof of MIVAR-Based Feature Selection Algorithm

In this section, the MIVAR-based feature selection algorithm is proven by ranking the aforementioned 70-dimensional all waveform features.

First, a network with a structure of $70 \times 35 \times 4$ is constructed and trained with all waveform features of the samples in the training set. To calculate the MIV of every dimension for different working conditions, we choose four samples that separately belong to NO, OR, IR, and RE and calculate the corresponding MIVs of every dimension using the algorithm proposed in Section 2, Step 2 to Step 5. Figures 4(a) to 4(d) show the MIVs of all 70-dimensional features for the four different states, where the x-coordinate represents the sequence number of the feature and the y-coordinate represents their MIVs.

In Figure 4, we separately mark three features with the largest MIVs among the 70 features of each state. Considering Figure 4(a) as an example, we mark the features numbers 33, 32, and 9 beside their columns with the form "$j(i)$," which means that the MIV of the ith feature places jth among the 70-dimensional features for NO. In other words, if we want to recognize a sample of NO with the network, these three features have the greatest effect on the recognition results. From Figure 4, we find that the sequence numbers of the top three features of different states are different, as listed in Table 1.

In Table 1, we can see that number 1 feature places first for IR, OR, and RE, thereby having the greatest effect on sample recognition, and number 33 feature places first for NO. It seems that features numbers 1, 2, and 3 are the best three features for classification because their MIVs are relatively larger than the MIVs of the other features in the three states. Correspondingly, feature number 33 is less suitable for classification because it performs well only for state NO. However, we claim that if a feature affects most of the states at the same level, as is the case with numbers 1, 2, and 3, it

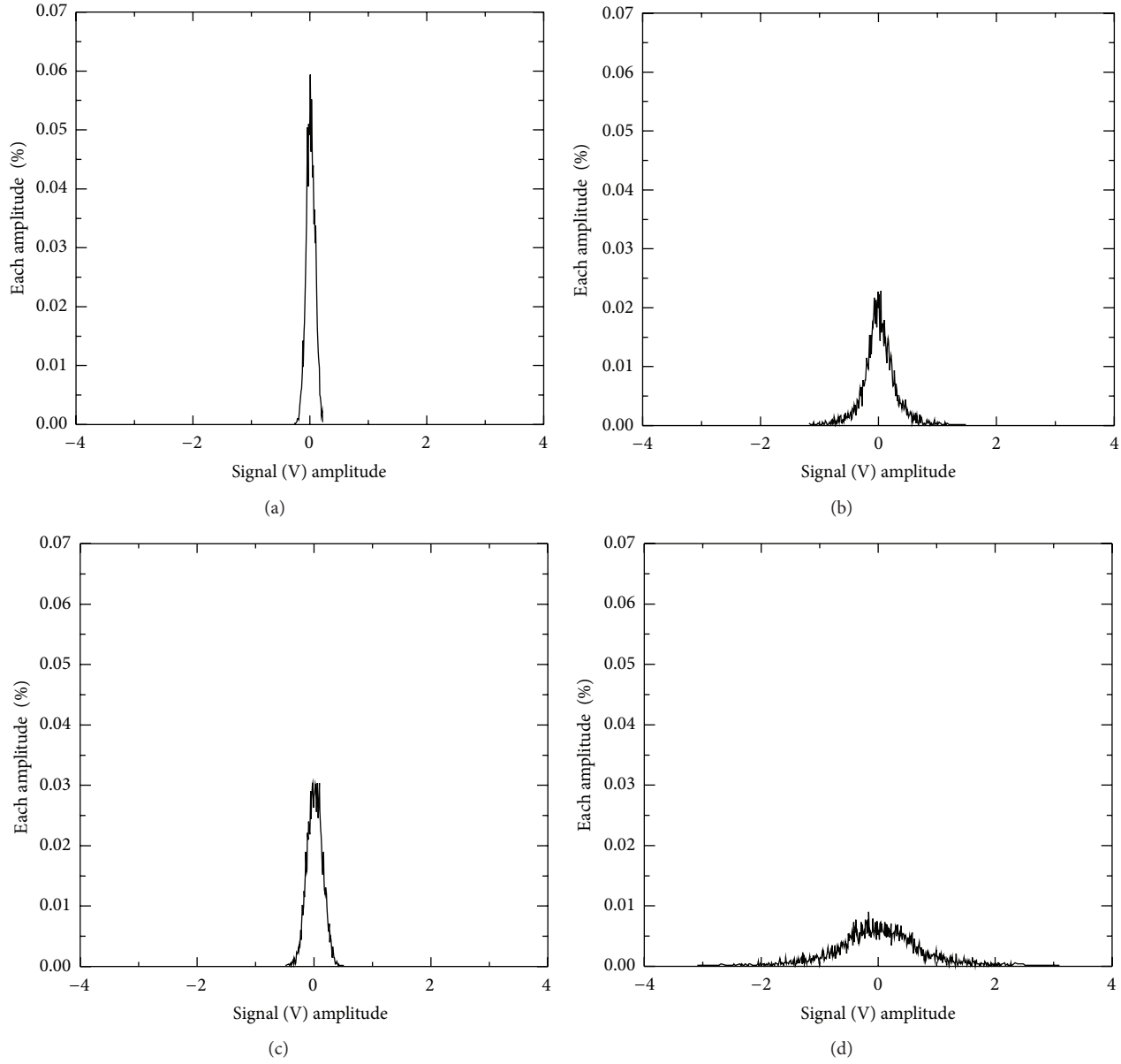


Figure 1: Probability density curves for four different running states: (a) NO, (b) IR, (c) RE, and (d) OR.

is probably not the most suitable feature for classification. At minimum, such a feature cannot efficiently classify the MIV types at the same level. On the contrary, feature number 33 might be more suitable for classification, even if its MIVs place first for NO only, because its MIVs for different classes are at different levels, which might make it a better feature for classification. Thus, the MIV cannot determine the feature that is more suitable for classification.

Second, the corresponding MIVARs of every feature are calculated by (5). Figure 5 shows the MIVARs of every feature in a histogram with the top ten features denoted in the form of "$j(i)$," which means that the MIVAR of the ith feature places jth among the 70-dimensional feature. The x-coordinate represents the feature sequence number, and the y-coordinate represents the MIVAR value.

In Figure 5, we can readily find the top ten features with the largest MIVARs among the 70-dimensional features. According to the MIVAR method, these ten features have the greatest effect on classification. We can see that the sequence numbers of these ten features are not the same as those in Table 2. As Figure 5 suggests, number 33 feature, whose MIVAR value places first among the 70 features, is the most efficient feature. However, it only performs well in just one state in Table 2. Only half of the features listed in Table 2 are marked in Figure 5; they are features numbers 1, 3, 9, 32, and 33. Among them, numbers 9, 32, and 33 perform well in only one state. According to the MIVAR method, numbers 33, 1, 32, 9, 3, 38, 4, 28, 35, and 19 are selected as the most efficient features for classification. The specific ranks of all 70-dimensional features are listed in Table 2.

Table 2: MIVAR ranks of 70-dimensional features.

Ranking number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sequence number	33	1	32	9	3	38	4	28	35	19
Ranking number	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Sequence number	40	10	2	11	23	7	47	12	6	31
Ranking number	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Sequence number	17	8	22	29	25	34	13	46	16	49
Ranking number	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
Sequence number	44	39	14	20	15	43	30	41	27	42
Ranking number	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**
Sequence number	24	50	58	18	21	5	57	37	45	52
Ranking number	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
Sequence number	26	54	55	48	53	36	51	59	61	62
Ranking number	**61**	**62**	**63**	**64**	**65**	**66**	**67**	**68**	**69**	**70**
Sequence number	56	60	63	64	66	67	65	68	70	69

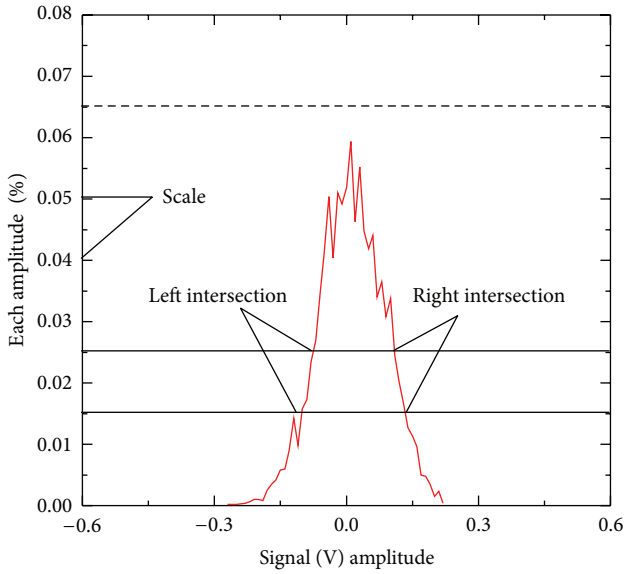


Figure 2: Schematic diagram for conceptual explanation.

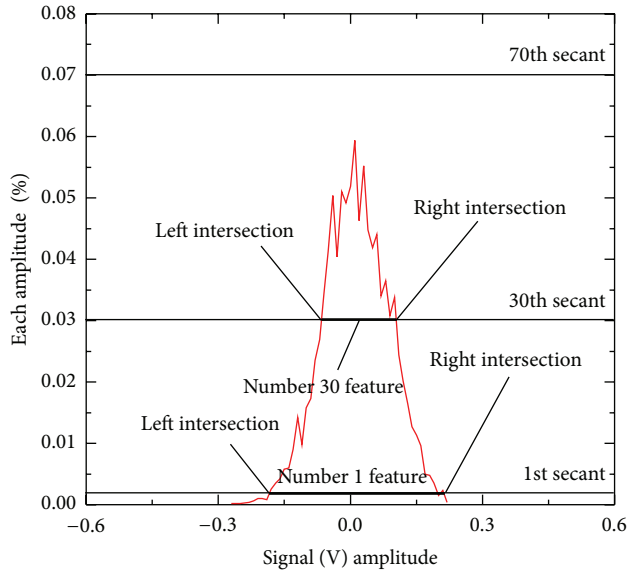


Figure 3: All waveform features with 70 dimensions.

Third, several comparisons are presented to prove the validity of the ranking results by constructing 14 groups of features as follows:

(1) Features 33, 1, 32, 9, 3, 38, 4, 28, 35, and 19, whose sequence numbers are the top ten;

(2) Features 40, 10, 2, 11, 23, 7, 47, 12, 6, and 31, whose sequence numbers are the second top ten;

(3) Features 17, 8, 22, 29, 25, 34, 13, 46, 16, and 49, whose sequence numbers are the third top ten;

(4) Features 44, 39, 14, 20, 15, 43, 30, 41, 27, and 42, whose sequence numbers are the fourth top ten;

(5) Features 24, 50, 58, 18, 21, 5, 57, 37, 45, and 52, whose sequence numbers are the fifth top ten;

(6) Features 26, 54, 55, 48, 53, 36, 51, 59, 61, and 62, whose sequence numbers are the sixth top ten;

(7) Features 56, 60, 63, 64, 66, 67, 65, 68, 70, and 69, whose sequence numbers are the bottom ten;

(8) new constructed 10-dimensional features with the top ten scores based on the PCA algorithm;

(9) new constructed 10-dimensional features with the second ten scores based on the PCA algorithm;

(10) new constructed 10-dimensional features with the third ten scores based on the PCA algorithm;

(11) new constructed 10-dimensional features with the fourth ten scores based on the PCA algorithm;

(12) new constructed 10-dimensional features with the fifth ten scores based on the PCA algorithm;

(13) new constructed 10-dimensional features with the sixth ten scores based on the PCA algorithm;

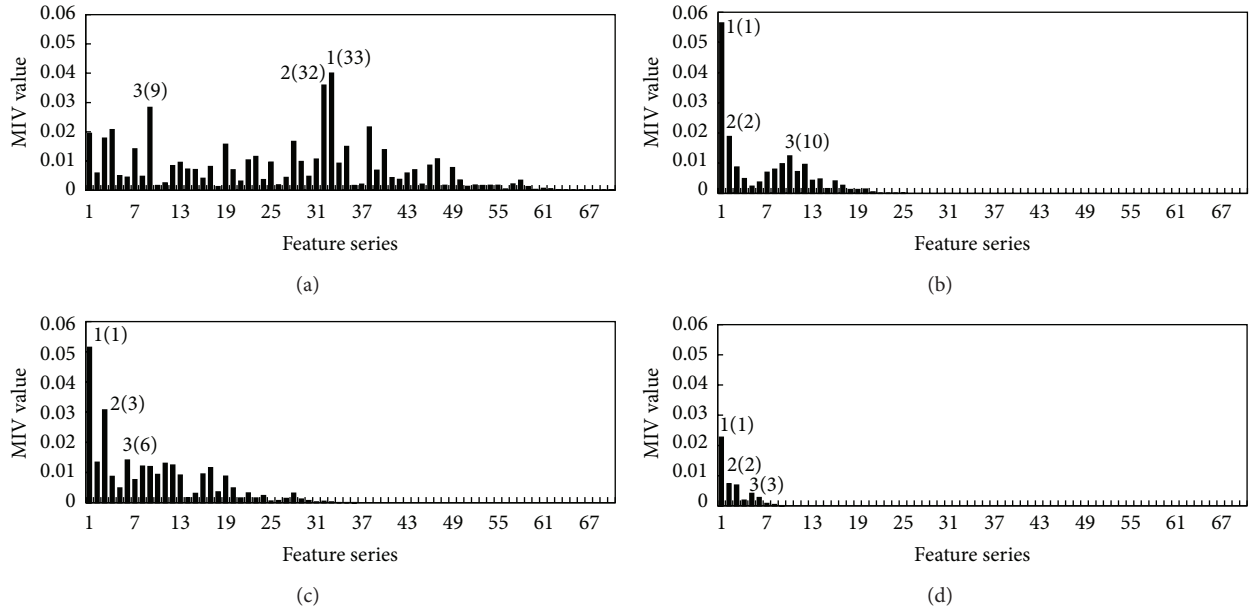


Figure 4: Feature MIVs in different states: (a) NO, (b) IR, (c) OR, and (d) RE.

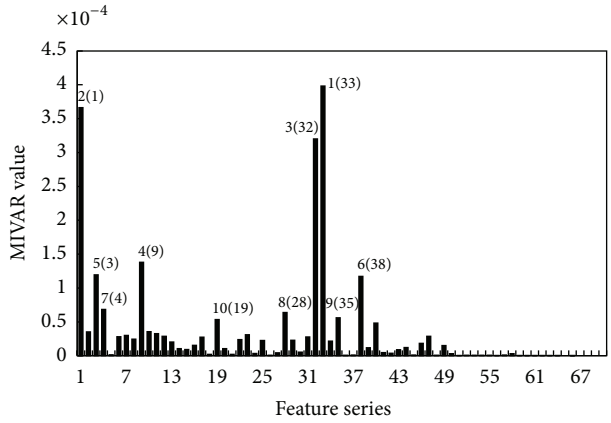


Figure 5: MIVAR histogram of every feature.

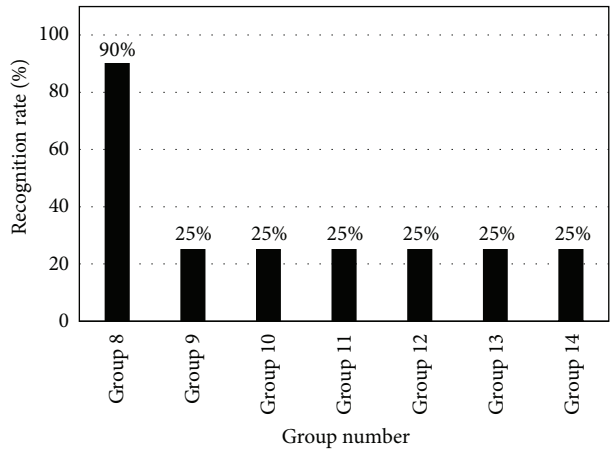


Figure 7: Recognition rate of PCA-based features.

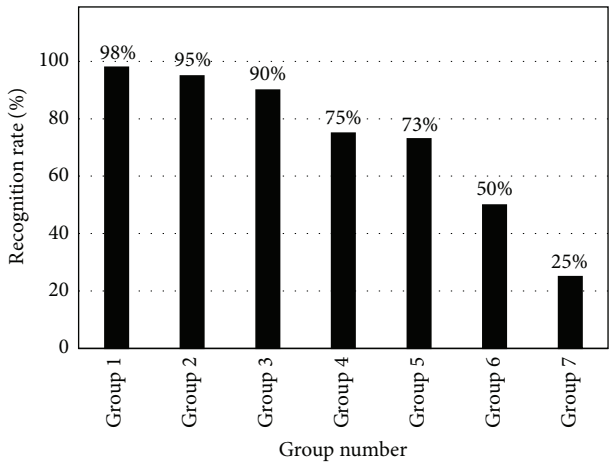


Figure 6: Recognition rate of MIVAR-based features.

(14) new constructed 10-dimensional features with the bottom ten scores based on the PCA algorithm.

In detail, 14 new training sets and testing sets are generated to train and test a newly constructed network with the structure 10×6×4. The corresponding recognition rates listed in Table 3 can be then used to prove whether the MIVAR-based ranking result is appropriate. To ensure the fairness of the comparison, the initial weights and training times during the training processes of the different groups should be the same.

According to the comparison results listed in Table 3, we can see that the features whose MIVAR ranking sequences are the top ten and the second top ten can lead to recognition rates of 98% and 95%, respectively. Moreover, the recognition rate decreases from 90% to 25% when the features in Groups 3 to 7 are used to represent the vibration signal. It is

Table 3: Recognition rate of different groups.

Network input	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
Recognition rate (%)	98	95	90	75	73	50	25
Network input	Group 8	Group 9	Group 10	Group 11	Group 12	Group 13	Group 14
Recognition rate (%)	90	25	25	25	25	25	25

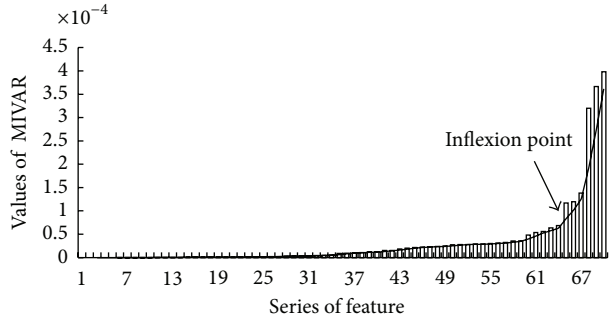


Figure 8: Ranking results of all 70-dimensional all waveform features.

proven that the MIVAR-based feature selection algorithm can be used to select more representative features from the multidimensional features. As for the features generated by the PCA algorithm, we use the features in Group 8 to Group 14 train of the same network. It is shown that the new constructed 10-dimensional features with the top ten scores can lead to the recognition rate of 90% and all the other 6 groups of 10-dimensional features can only lead to 25% recognition rate. Figures 6 and 7 show their histograms. By comparing the recognition rates in Figures 6 and 7, we find that the recognition rate of Group 8 is not as good as the ones of Group 1 and 2 which can partly explain the advantage of MIVAR based feature selection algorithm. It should be mentioned that the principal component contribution rates summation of the top 3 vectors is more than 95%. So, the features in Group 9 to Group 14 are useless for the final classification.

Last, we display the 70-dimensional all waveform features in the order of the corresponding MIVAR value in Figure 8, where the MIVAR of all the features are displayed by hollow histograms, and the corresponding trend line is presented simultaneously. We can see that the MIVAR value of the 65th histogram indicated by the arrow is obviously larger than for the 64th and the changing rate of the trend line after the 65th feature is much larger than before it. This way, we consider 65th feature as the inflexion point and recommend the features (number 33, 1, 32, 9, 3, and 38) whose MIVAR values are the top six largest most representative features when the BP network is used as the classifier.

5. Conclusion

In this paper, a MIVAR method was proposed to determine the feature that is more suitable for ANN-based classification. The MIVAR values of all the features were calculated by changing the input vectors and then measuring the differences of the output vectors after the training process of the BP network. It was proven that using the features with higher MIVAR values can lead to higher recognition rates.

As an example, 70-dimensional all waveform features of a rolling bearing vibration signal were ranked based on the MIVAR method. The features with the largest ten MIVAR values can lead to a recognition rate of 98%, and the corresponding recognition rate of the second, third, fourth, fifth, sixth, and seventh largest ten MIVAR values are 95%, 90%, 75%, 73%, 50%, and 25%, respectively. This decreased recognition rate proved the effectiveness of the MIVAR method. To compare the effectiveness of the MIVAR method to the traditional algorithm, the PCA algorithm is then used to generate 7 groups of 10-dimensional features (Group 8 to Group 14). And the 10-dimensional features with the top ten scores can lead to a recognition rate of 90%, which is not as good as that for Groups 1 and 2.

In addition, it should be pointed out that the discussion is limited to the use of time-domain features to describe a steady vibration signal. Moreover, the MIVAR algorithm can be extended also to the selection of frequency-domain features.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (51275030) and the "Fundamental Research Funds for the Central Universities M11JB00210."

References

[1] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[2] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *The Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[3] H. Brunzell and J. Eriksson, "Feature reduction for classification of multidimensional data," *Pattern Recognition*, vol. 33, no. 10, pp. 1741–1748, 2000.

[4] S. B. Serpico, M. D'Inca, F. Melgani, and G. Moser, "Comparison of feature reduction techniques for classification of hyperspectral remote sensing data," in *Image and Signal Processing for Remote Sensing VIII, 347*, vol. 4885 of *Proceedings of SPIE*,

International Society for Optics and Photonics, Crete, Greece, September 2002.

[5] G. Isabelle and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[6] H. K. Ekenel and B. Sankur, "Feature selection in the independent component subspace for face recognition," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1377–1388, 2004.

[7] A. Malhi and R. X. Gao, "PCA-based feature selection scheme for machine defect classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 6, pp. 1517–1525, 2004.

[8] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, "Feature selection using principal feature analysis," in *Proceedings of the 15th International Conference on Multimedia*, New York, NY, USA, 2007.

[9] J. Utans, J. Moody, S. Rehfuss, and H. Siegelmann, "Input variable selection for neural networks: application to predicting the U.S. business cycle," in *Proceedings of the IEEE/IAFE Computational Intelligence for Financial Engineering (CIFEr '95)*, pp. 118–122, New York, NY, USA, April 1995.

[10] C. Huang and C. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," *Expert Systems with Applications*, vol. 31, no. 2, pp. 231–240, 2006.

[11] B. Samanta, K. R. Al-Balushi, and S. A. Al-Araimi, "Artificial neural networks and genetic algorithm for bearing fault detection," *Soft Computing*, vol. 10, no. 3, pp. 264–271, 2006.

[12] X. Zhang, X. Lu, Q. Shi et al., "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data," *BMC Bioinformatics*, vol. 7, article 197, 2006.

[13] B. S. Yang, T. Han, and J. L. An, "ART-KOHONEN neural network for fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 18, no. 3, pp. 645–657, 2004.

[14] Y. Lei, Z. He, Y. Zi, and Q. Hu, "Fault diagnosis of rotating machinery based on multiple ANFIS combination with GAs," *Mechanical Systems and Signal Processing*, vol. 21, no. 5, pp. 2280–2294, 2007.

[15] http://csegroups.case.edu/bearingdatacenter/home.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 590787, 11 pages
http://dx.doi.org/10.1155/2014/590787

Research Article

Strain Rate Dependent Deformation of a Polymer Matrix Composite with Different Microstructures Subjected to Off-Axis Loading

Xiaojun Zhu, Xuefeng Chen, Zhi Zhai, Zhibo Yang, Xiang Li, and Zhengjia He

State Key Lab for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Correspondence should be addressed to Xuefeng Chen; chenxf@mail.xjtu.edu.cn

Received 24 April 2014; Accepted 2 June 2014; Published 23 June 2014

Academic Editor: Weihua Li

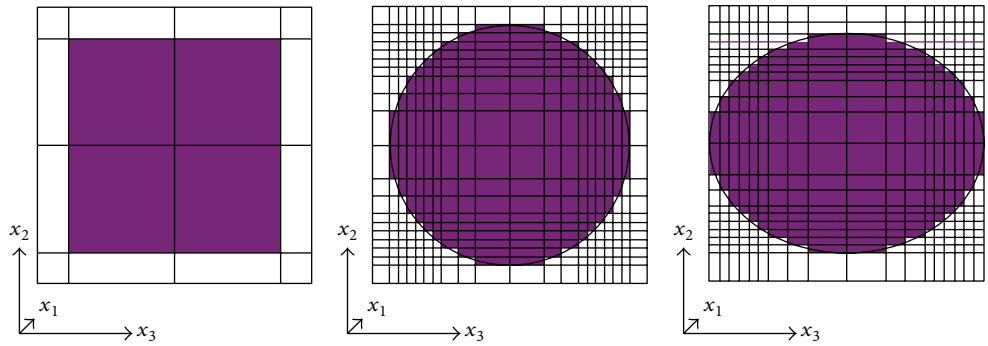
This paper aims to investigate the comprehensive influence of three microstructure parameters (fiber cross-section shape, fiber volume fraction, and fiber off-axis orientation) and strain rate on the macroscopic property of a polymer matrix composite. During the analysis, AS4 fibers are considered as elastic solids, while the surrounding PEEK resin matrix exhibiting rate sensitivities are described using the modified Ramaswamy-Stouffer viscoplastic state variable model. The micromechanical method based on generalized model of cells has been used to analyze the representative volume element of composites. An acceptable agreement is observed between the model predictions and experimental results found in the literature. The research results show that the stress-strain curves are sensitive to the strain rate and the microstructure parameters play an important role in the behavior of polymer matrix.

1. Introduction

In the last few decades, polymer matrix composite materials (PMCs) have been developed rapidly to meet the demands for better materials with higher standards of performance and reliability in structures and machines [1, 2]. In some of these applications such as marine structures, aerospace, and lightweight armor, the PMCs are often subjected to complex loadings under extreme circumstances [3, 4] in which the properties of the PMCs exhibit highly nonlinear and rate dependence, so it is necessary for structural design and analysis to characterize and model the nonlinearity and strain rate dependence of the composite.

Polymers are known to have a strain rate dependent deformation response that is nonlinear above 1 or 2% strain [5]. Many experimental studies have been made to determine the effects of strain rate on the PMCs [6]. Weeks [7] conducted experiments using an MTS machine and the split Hopkinson pressure bar for AS4/PEEK composite and produced strain rates ranging from 0.00001/s to 1000/s. Uniaxial tension tests were conducted on various off-axis coupon specimens to obtain stress/strain curves for various strain rates [8]. Haque and Ali [9] adopted a systematic experimental approach to identify the damage progression at various stress levels and the strain rate effects on composites. Shokrieh and Omidi [10] studied tensile failure properties unidirectional glass/epoxy composites at various strain rates from 0.001/s to 100/s using a high-speed servohydraulic testing apparatus. Experimental results showed a significant increase of the tensile strength by increasing the strain rate.

On the other hand, there are also many macromechanical and micromechanical models to predict the behavior of composite materials subjected to different strain rates [11, 12]. Weeks and Sun [13] developed a macromechanical, rate dependent constitutive model to analyze the inelastic response of carbon reinforced composites. Thiruppukuzhi and Sun [14] later directly incorporated the rate dependence of the material response into the constitutive model. Espinosa et al. [15] presented a 3D finite deformation anisotropic viscoplasticity model to analyze the effects of strain rate and temperature on a woven composite made of S-2 glass fibers. A 3D model based on finite elastoplasticity was applied to study the effect of temperature and strain rate on the tensile behaviour on a series of polymeric matrix unidirectional glass-fibre composites [16]. Recently, a Johnson-Cook based



(a) Square (b) Circular (c) Elliptical

Figure 1: Three kinds of fiber shapes.

modeling approach was used to represent the apparent strain rate dependency of textile reinforced composites in laminate through-thickness direction [17]. A phenomenological-based approach was proposed by Raimondo et al. [18] for the three-dimensional modeling of strain rate in unidirectional polymer composites. A nonlinear constitutive model for large deformation loading at different strain rate condition was developed to represent tensile progressive damage of the nonlinear large deformation rate dependent behavior of polymer-based composite materials [19].

Compared with macromechanical model, which considered composites as anisotropic medium with homogeneous distribution, the micromechanical model only needs to test the ingredient properties of composites, while macromechanical model needs to do repetitive experiments for composites [20]. Therefore, many scholars have done a lot of research on the micromechanical model for years. A 3D micromechanical formulation was proposed [21] for fiber composites with viscoplastic matrix properties. The nonlinear responses of composites under various cyclic loading conditions were predicted accurately by their analysis. Goldberg and Stouffer [22] adopted a four-region micromechanics method, in which the composite unit cell is divided into a number of slices to analyze polymer matrix composites subject to different strain rates. Later, the micromechanical model was implemented in the nonlinear finite element software LS-DYNA [23]. By combining the bridging micromechanics model [24] with classical lamination, a general constitutive relationship was established for the inelastic and failure analysis of laminate structures [25]. Paley and Aboudi [26] proposed the generalized method of cells (GMC) to deal with the representative volume element (RVE) with complex microstructures. Ogihara et al. [27] adopted the GMC to study the nonlinear behavior of unidirectional carbon-epoxy laminates subjected to off-axis loading. The epoxy matrix was predicted using the one-parameter plasticity model. Tsai and Chen [28] employed the GMC to characterize the nonlinear rate-dependent behaviors of graphite/epoxy composites. The epoxy matrix is described by a three-parameter viscoplasticity model. However, the comprehensive effect of three microstructure parameters (fiber cross-section shape, fiber volume fraction, and fiber off-axis orientation) and the strain rate on the macroscopic property of composites has been seldom reported in the above studies. In this paper, by combining the GMC with the modified Ramaswamy-Stouffer viscoplastic state variable model, and with no need to judge whether the material is in elastic or plastic stage and is more convenience and effective to predict the matrix behavior [29], a new general constitutive relationship was established for the inelastic analysis of the comprehensive influence of three microstructure parameters and strain rate on the stress-strain behavior of the polymer matrix composite.

In this paper, the rest outline is as follows. Section 2 introduces the micromechanical model based GMC. In Section 3, the modified Ramaswamy-Stouffer viscoplastic state variable model is incorporated into GMC. Composites with three microstructure parameters are considered to analyze the rate dependent stress-strain response in Section 4. Conclusions are given in Section 5.

2. Micromechanical Model Based Generalized Model of Cells

2.1. Generalized Model of Cells. The two-dimensional generalized method of cells is a micromechanical model developed originally by Paley and Aboudi [26] for predicting the response of unidirectional matrix composites with periodic microstructures. The GMC was then reformulated in terms of the interfacial subcell tractions substituting the subcell strains as the basic unknowns by Pindera and Bednarcyk [30], which can significantly increase the calculation efficiency when the number of subcells became larger.

When a micromechanical approach is used to model the mechanical response of fiber reinforced composites with periodic microstructures, a proper RVE is required to represent the microstructures of the materials such that the overall composites responses can be predicted directly from the representative volume element. In this study, three kinds of fiber cross-section shapes, such as square, circular, and elliptical, were considered as shown in Figure 1. In this figure, the fiber and matrix are indicated by the black and white, respectively.

In the GMC analysis, the representative volume element is usually divided into $N_\beta \times N_\gamma$ subcells as shown in Figure 2.

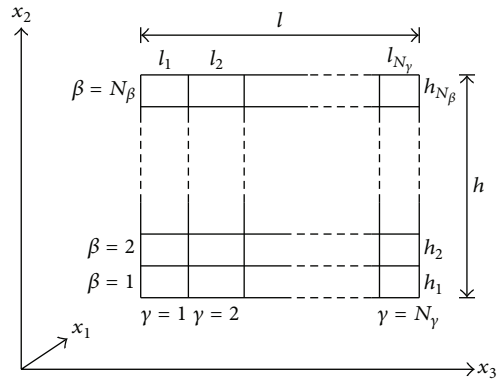


FIGURE 2: A typical RVE divided into $N_\beta \times N_\gamma$ subcells in GMC analysis.

In general, each of these subcells is assumed to be occupied by a material that exhibits inelastic behavior. The subcell material's inelastic behavior can be modeled by a lot of constitutive theories, such as classical incremental plasticity, linear viscoelasticity, or unified viscoplasticity theories. Therefore, the representative volume element, which consists of $N_\beta \times N_\gamma$ different inelastic materials, can represent a multiphased, inelastic composite.

Based on the displacement continuity on the interface of the adjacent subcells in conjunction with the periodicity condition of the RVE, the relation between overall strain and the subcell strain is expressed as

$$\overline{\boldsymbol{\varepsilon}}^{(\beta\gamma)} = \mathbf{A}^{(\beta\gamma)}\overline{\boldsymbol{\varepsilon}}, \tag{1}$$

where $(\beta\gamma)$ indicates the subcell whose location in RVE is at the βth row and γth column, and $\mathbf{A}$ is the matrix linking the micro- and macrostrain.

At the same time, the interfacial traction continuity conditions can be expressed as

$$\begin{aligned}
&\overline{\sigma}_{22}^{(1\gamma)} = \overline{\sigma}_{22}^{(2\gamma)} = \cdots = \overline{\sigma}_{22}^{(N_\beta\gamma)} = \mathbf{T}_{22}^{(\gamma)} \quad \left(\gamma = 1, \ldots, N_\gamma\right),\\
&\overline{\sigma}_{33}^{(\beta 1)} = \overline{\sigma}_{33}^{(\beta 2)} = \cdots = \overline{\sigma}_{33}^{(\beta N_\gamma)} = \mathbf{T}_{33}^{(\beta)} \quad \left(\beta = 1, \ldots, N_\beta\right),\\
&\overline{\sigma}_{21}^{(1\gamma)} = \overline{\sigma}_{21}^{(2\gamma)} = \cdots = \overline{\sigma}_{21}^{(N_\beta\gamma)} = \mathbf{T}_{21}^{(\gamma)} = \mathbf{T}_{12}^{(\gamma)}\\
&\qquad\qquad \left(\gamma = 1, \ldots, N_\gamma\right),\\
&\overline{\sigma}_{31}^{(\beta 1)} = \overline{\sigma}_{31}^{(\beta 2)} = \cdots = \overline{\sigma}_{31}^{(\beta N_\gamma)} = \mathbf{T}_{31}^{(\beta)} = \mathbf{T}_{13}^{(\beta)}\\
&\qquad\qquad \left(\beta = 1, \ldots, N_\beta\right),\\
&\overline{\sigma}_{23}^{(1\gamma)} = \overline{\sigma}_{23}^{(2\gamma)} = \cdots = \overline{\sigma}_{23}^{(N_\beta\gamma)} = \mathbf{T}_{23}^{(\gamma)} = \mathbf{T}_{23}\\
&\qquad\qquad \left(\gamma = 1, \ldots, N_\gamma\right),\\
&\overline{\sigma}_{32}^{(\beta 1)} = \overline{\sigma}_{32}^{(\beta 2)} = \cdots = \overline{\sigma}_{32}^{(\beta N_\gamma)} = \mathbf{T}_{32}^{(\beta)} = \mathbf{T}_{23}\\
&\qquad\qquad \left(\beta = 1, \ldots, N_\beta\right).
\end{aligned} \tag{2}$$

For each subcell of composites, the constitutive relationship of each subcell can be written as

$$\overline{\boldsymbol{\varepsilon}}^{(\beta\gamma)} = \mathbf{S}^{(\beta\gamma)}\overline{\sigma}^{(\beta\gamma)} + \overline{\boldsymbol{\varepsilon}}^{p(\beta\gamma)} + \boldsymbol{\alpha}^{(\beta\gamma)}\Delta T. \tag{3}$$

Substituting (3) into (1) and then combining (2), the relations between subcell tractions and overall strains can be obtained as

$$\overline{\sigma}_{ij}^{(\beta\gamma)} = \mathbf{C}_{ijkl}^{(\beta\gamma)}\mathbf{A}^{(\beta\gamma)}\overline{\boldsymbol{\varepsilon}}. \tag{4}$$

Based on the homogenization theory, the overall stress of the RVE can be written as

$$\overline{\boldsymbol{\sigma}} = \frac{1}{hl}\sum_{\beta=1}^{N_\beta}\sum_{\gamma=1}^{N_\gamma} h_\beta l_\gamma \overline{\boldsymbol{\sigma}}^{(\beta\gamma)}. \tag{5}$$

Substituting (4) into (5), the overall stress and strain relation of the RVE are established as

$$\overline{\boldsymbol{\sigma}} = \mathbf{C}^*\left(\overline{\boldsymbol{\varepsilon}} - \overline{\boldsymbol{\varepsilon}}^p - \boldsymbol{\alpha}^*\Delta T\right), \tag{6}$$

where $\mathbf{C}^*$ indicates the overall elastic stiffness matrix, $\overline{\boldsymbol{\varepsilon}}^p = [\overline{\varepsilon}_{11}^p, \overline{\varepsilon}_{22}^p, \overline{\varepsilon}_{33}^p, \overline{\varepsilon}_{23}^p, \overline{\varepsilon}_{13}^p, \overline{\varepsilon}_{12}^p]^T$ indicates the overall plastic strain, and $\boldsymbol{\alpha}^* = [\alpha_{11}^*, \alpha_{22}^*, \alpha_{33}^*]^T$ represents the overall thermal expansion coefficient vector.

It should be noted that the elements of matrixes $\mathbf{C}^*$, $\overline{\boldsymbol{\varepsilon}}^p$, and $\boldsymbol{\alpha}^*$ can be explicitly obtained in terms of the subcell material and geometric parameters and subcell plastic strains, so when the subcell ingredient properties and the RVE geometry are known, (6) can be used to model the responses of fiber composites.

3. Viscoplastic Constitutive Model

The matrix viscoplastic constitutive model is based on the modified Ramaswamy-Stouffer viscoplastic state variable model. The Ramaswamy-Stouffer viscoplastic state variable model [31] was originally developed to simulate the rate dependent inelastic response of metals. However, the relationship between load and deformation in resins is more complicated than that in metals since the hydrostatic component of the stress has a significant effect even at low level of stress [32]. The effect of the hydrostatic stresses was considered by modifying the effective stress term in the flow law of Ramaswamy-Stouffer model [22]. In the modified Ramaswamy-Stouffer model, the total strain rate, $\dot{\varepsilon}_{ij}$, is composed of elastic strain rate, $\dot{\varepsilon}_{ij}^e$, and inelastic strain rate, $\dot{\varepsilon}_{ij}^I$; that is,

$$\dot{\varepsilon}_{ij} = \dot{\varepsilon}_{ij}^e + \dot{\varepsilon}_{ij}^I. \tag{7}$$

The elastic strain rate can be obtained according to the time derivative of Hook's law. The inelastic strain rate is defined in the following form:

$$\dot{\varepsilon}_{ij}^I = D_0 \exp\left[-\frac{1}{2}\left(\frac{Z_0^2}{3K_2}\right)^n\right] \times \frac{s_{ij} - \Omega_{ij}}{\sqrt{K_2}}, \tag{8}$$

where D_0, Z_0, and n are all material constants. D_0 denotes the maximum inelastic strain rate, Z_0 indicates the initial, isotropic "hardness" of the material before any load is applied, n represents the rate dependence of deformation response, S_{ij} is the deviatoric stress component, and Ω_{ij} is the internal stress state variable.

The relation between the internal stress rate, $\dot{\Omega}_{ij}$ and Ω_{ij}, is defined as follows:

$$\dot{\Omega}_{ij} = \frac{2}{3} q \Omega_m \dot{\varepsilon}_{ij} - q \Omega_{ij} \dot{\varepsilon}_e^I, \tag{9}$$

where q and Ω_m are both material constants. q represents the "hardening" rate, Ω_m represents the maximum value of the internal stress, and $\dot{\varepsilon}_e^I$ is the effective inelastic strain rate, which is defined as follows:

$$\dot{\varepsilon}_{ij}^I = \sqrt{\frac{2}{3} \dot{\varepsilon}_{ij} \dot{\varepsilon}_{ij}}. \tag{10}$$

The term K_2, which represents the effective stress, is defined in the original Ramaswamy-Stouffer model in the following form:

$$K_2 = \frac{1}{2} \left(S_{ij} - \Omega_{ij} \right) \left(S_{ij} - \Omega_{ij} \right). \tag{11}$$

In the modified Ramaswamy-Stouffer model, in order to consider the effect of hydrostatic stresses, (11) is rewritten as follows:

$$K_2 = \frac{1}{2} \left[K_{11} + K_{22} + K_{33} + 2 \left(K_{12} + K_{13} + K_{23} \right) \right]. \tag{12}$$

The normal terms in the above expression are the same as the original definition while the shear terms are modified and can be written as

$$\begin{aligned} K_{12} &= \alpha \left(S_{12} - \Omega_{12} \right) \left(S_{12} - \Omega_{12} \right), \\ K_{13} &= \alpha \left(S_{13} - \Omega_{13} \right) \left(S_{13} - \Omega_{13} \right), \\ K_{23} &= \alpha \left(S_{23} - \Omega_{23} \right) \left(S_{23} - \Omega_{23} \right), \end{aligned} \tag{13}$$

where

$$\alpha = \left(\frac{\sigma_m}{\sqrt{J_2}} \right)^{\beta}. \tag{14}$$

In (14), σ_m is the mean stress, J_2 is the second invariant of the deviatoric stress tensor, and β is a rate independent material constant which is determined empirically by fitting data from uniaxial composites with shear dominated fiber orientation angles, such as [15°]. The other material constants, such as D_0, Z_0, and n, are determined through the method discussed in the article written by Goldberg and Stouffer [22].

Through the above introduction of the modified Ramaswamy-Stouffer model, it can be seen that the model does not depend on the yield rule and the inelastic strains are assumed to be present at all values of stress. Therefore, there is no need to judge whether the material is in elastic or plastic stage.

Table 1: Material properties of PEEK resin [7].

E (GPa)	ν	D_0 (1/sec)	n	Z_0 (MPa)	q	Ω_m (MPa)	β
4.0	0.4	10^4	0.7	630	310	52	0.40

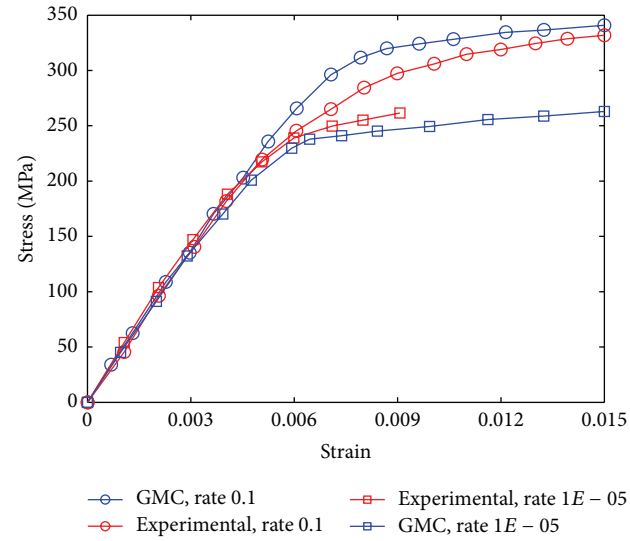


Figure 3: Stress-strain response of AS4/PEEK [15°] laminate at strain rate of 0.1/sec and 10^{-5}/sec.

4. Results and Discussion

4.1. Model Validation. To verify the ability of the micromechanics model and the viscoplastic constitutive model in the prediction of rate effects of composites several examples are considered and discussed in this section. The material considered here is a composite composed of carbon AS4 fibers in a PEEK thermoplastic matrix. For the AS4 fibers, the longitudinal elastic modulus is 214 GPa, the transverse and in-plane shear modulus is 14 GPa, the longitudinal Poisson's ratio is 0.2, and the transverse Poisson's ratio is 0.25 [22]. The material properties of PEEK resin can be seen in Table 1. The fiber volume fraction (v_f) used here is 0.62 and the fiber cross-section shape is square (seen in Figure 1). For comparison purposes, the experimental data obtained by Weeks [7] is shown as well. Two different strain rates, 0.1/sec and 10^{-5}/sec (which is written as 1E-05 in the figures for convenience), are considered. From Figures 3, 4, and 5, it can be seen that the results predicted by the presented micromechanics model and viscoplastic constitutive model exhibit good agreement with the experimental results.

4.2. Stress-Strain Response of Composites with the Same Fiber Volume Fraction but Different Fiber Shapes and Different Strain Rates in Different Fiber Off-Axis Orientations. Figure 6 presents the responses for 15°, 45°, 60°, and 75° off-axis orientations in the case that composites contain 0.15 fiber volume fraction with different fiber shapes and strain rates. In the case of the elliptical fibers, the transverse loading is applied in the principal material directions of the long axis. In this kind of composites with very low fiber volume fraction,

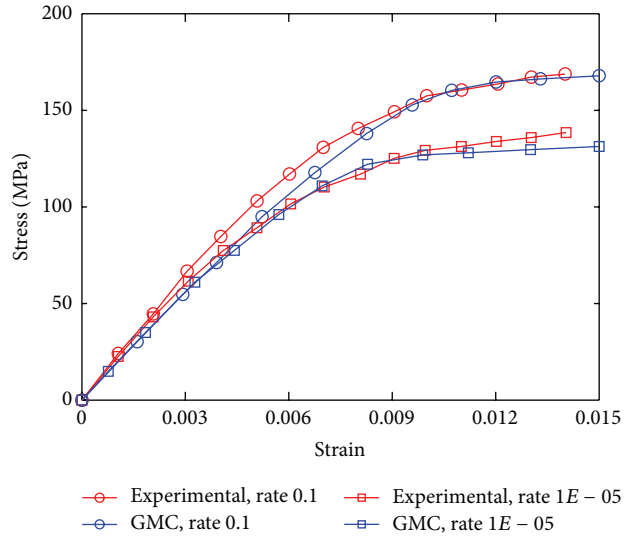


Figure 4: Stress-strain response of AS4/PEEK [30°] laminate at strain rate of 0.1/sec and 10^{-5}/sec.

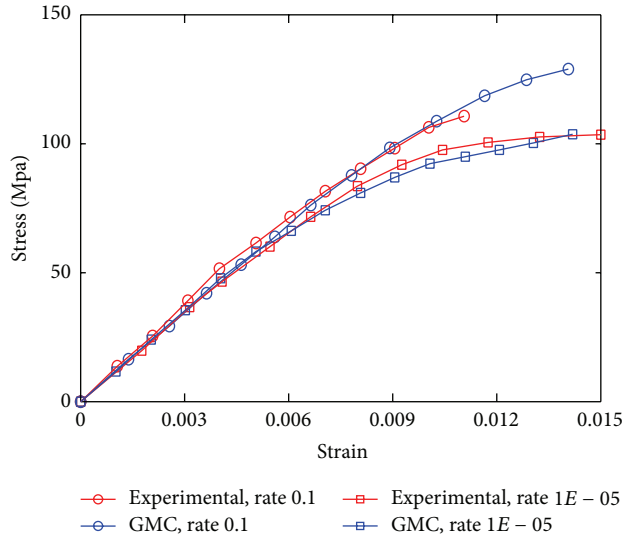


Figure 5: Stress-strain response of AS4/PEEK [45°] laminate at strain rate of 0.1/sec and 10^{-5}/sec.

the effect of fiber on the composites behavior is small, so it can be seen that the composites response is hardly affected by the fiber cross-section shape, but it could be affected by the off-axis orientation and the stain rate. Among the four kinds of off-axis orientations, the one with 15° off-axis orientation exhibits the stiffest response while the one with 60° off-axis orientation exhibits the most compliant response. For all the off-axis orientations, when the strain rate changes from 10^{-5}/sec to 0.1/sec, the composites provide an effective increase in the flow stress while the elastic behavior almost remain unchanged. This is because the fact that when the strain rate is smaller, the composites have more time to occur plastic flow and unload.

Increasing the fiber volume fraction further accentuates the differences in the composite's transverse response due to the fiber's cross-sectional shape. Figure 7 shows the stress-strain responses of composites when the fiber volume fraction is increased to 0.30. It can be seen that when the off-axis angle is smaller than 75°, the composites response is hardly affected by the fiber cross-section shape. However, when the off-axis angle is increased to 75°, the effect of the fiber cross-sectional shape on the transverse response in the plastic region becomes discernible, with the square fibers being the most effective in increasing the flow stress of the composite. Figure 7(d) shows that the responses of composites with circular fibers and elliptical fibers with an aspect ratio of 4/3 are almost the same, which are lower than the responses of composites with square fibers. With the increasing of the off-axis orientations of composites, the response of composites decreases first and then increases.

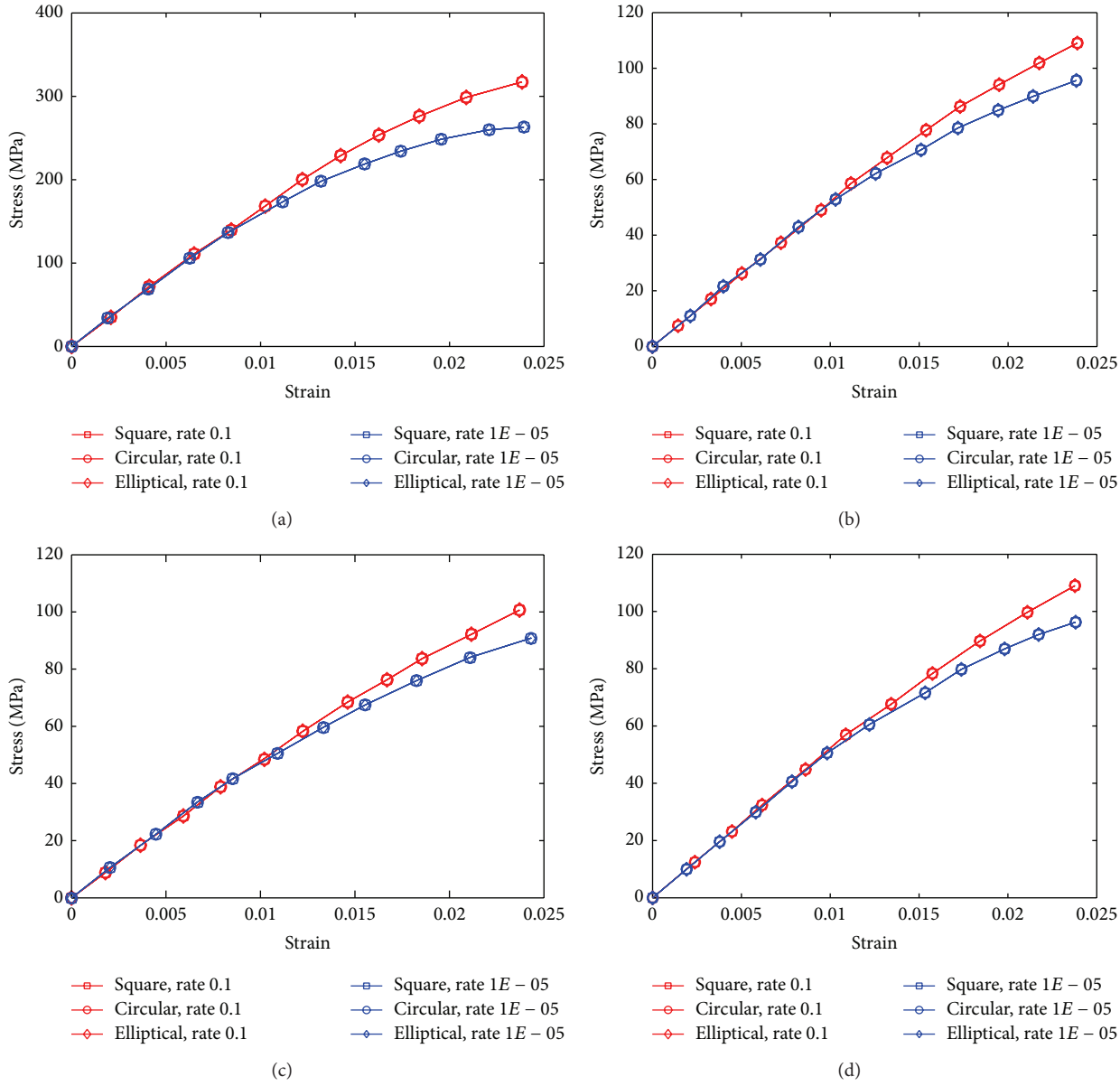


FIGURE 6: Off-axis responses of AS4/PEEK laminate ($v_f = 0.15$) at strain rate of 0.1/sec and 10^{-5}/sec: (a) 15°, (b) 45°, (c) 60°, and (d) 75°.

Figure 8 presents the results that correspond to those shown in the preceding two figures when the fiber volume fraction is further increased to 0.45. In this case, compared with the stress-strain curves when the fiber volume fraction is 0.30, it can be seen that when the off-axis angle is 60°, the composites response has already been affected by the fiber cross-section shape although the difference is small. But when the off-axis angle is increased to 75°, a substantial difference between the unit cell with the square fiber and the remaining unit cells is now apparent in the plastic region. In Figure 8(d), the square fiber provides a 20% increase in the transverse flow stress of the composite relative to that of the elliptical and circular fibers when the strain rate is 10^{-5}/sec, while the square fiber provides a 10% increase when the strain rate is 0.1/sec. This is because the square fiber can provide a higher magnitude of hydrostatic stress in the matrix phase relative to the circular fiber, which can delay localized yielding and provide constraint on the expansion of the plastic zone throughout the matrix phase. When the strain rate is smaller, the composites have more time to occur plastic strain and unload. Therefore, it can be noted that when the strain rate is 10^{-5}/sec, the difference of composites response between the square fiber and the circular fiber is larger than the case when the strain rate is 0.1/sec.

Figure 9 shows the stress-strain curves when the fiber volume fraction is increased to 0.55. This fiber volume fraction is close to the maximum allowable for the RVE with the elliptical fiber, which is limited by the contact of fibers along the major axis in two adjacent RVE. This contact occurs when the fiber volume fraction is 0.59 in the case of fibers

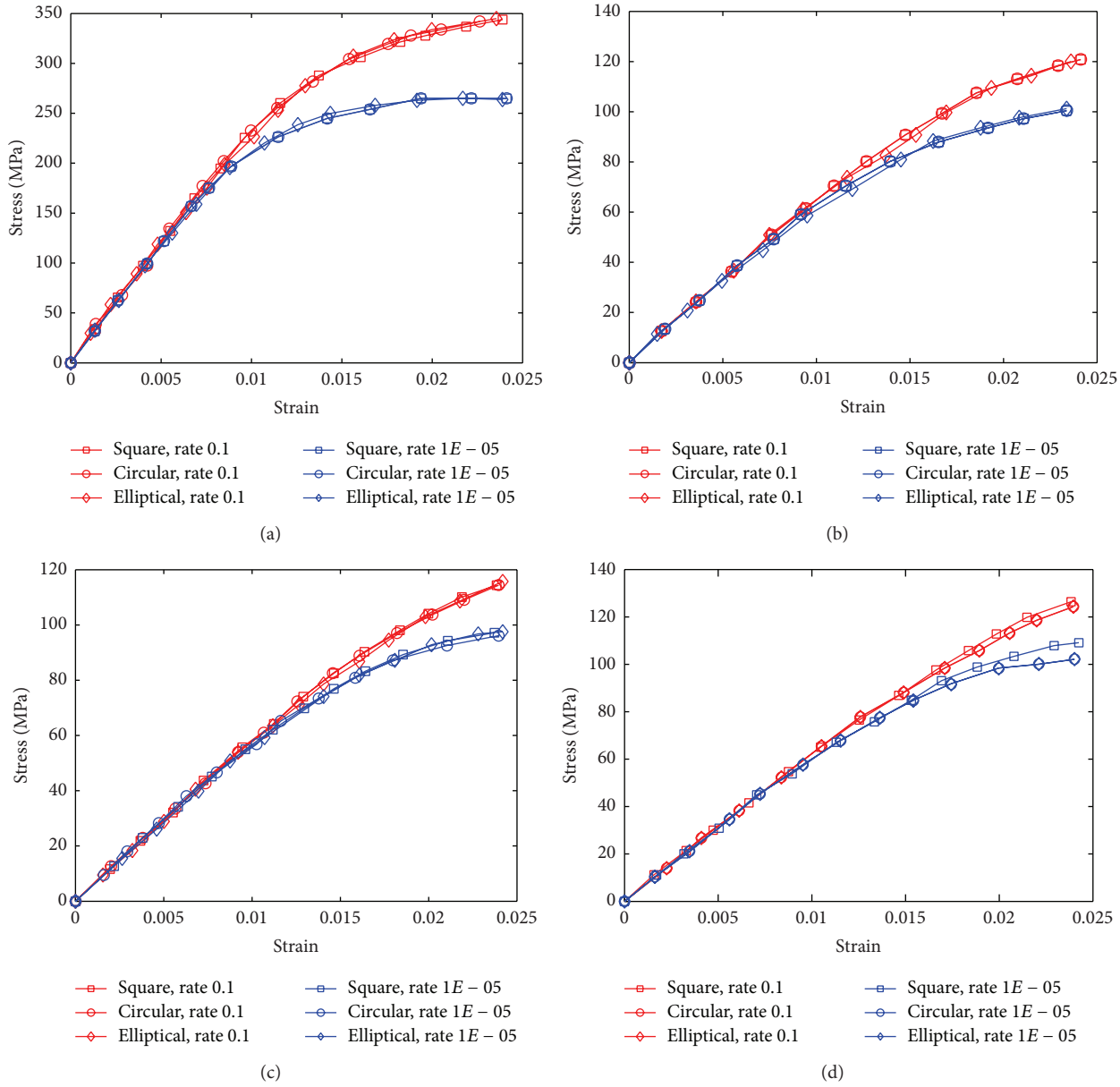


Figure 7: Off-axis responses of AS4/PEEK laminate ($v_f = 0.30$) at strain rate of 0.1/sec and 10^{-5}/sec: (a) 15°, (b) 45°, (c) 60°, and (d) 75°.

with an aspect ratio of 4/3. From Figure 9(b), it can be seen that the composites response is affected by the fiber cross-section shape when the off-axis angle is just 45°, which is smaller than the preceding cases. In additionally, for both of the two kinds of strain rates, the difference of the three kinds of fibers is more obvious with the increase of the off-axis angle. In Figure 9(d), the difference between the composites with square fibers and the composites with circular fibers is very big, and the response of the composites with square fibers at the strain rate of 10^{-5}/sec is almost the same as the response of the composites with circular fibers at the strain rate of 0.1/sec.

4.3. Stress-Strain Response of Composites with the Same Fiber Off-Axis Orientation but Different Fiber Shapes and Fiber Volume Fractions at Different Strain Rates. Figure 10 shows the stress-strain response for 0.15, 0.30, 0.45, and 0.55 fiber volume fractions in the case that composites fiber off-axis angle is 90°. It can be seen that when the fiber volume fraction is less than 0.30, the stress-strain response is barely affected by the fiber cross-section shapes. When the fiber volume fraction is more than 0.30, the difference between different fiber cross-section shapes can be obtained. With the increase of the fiber volume fraction, the difference becomes larger and the stiffness of composites will increase, which is due to the bigger stiffness of fiber. When the fiber volume is increased to 0.45, the response of the composites with square fibers at the strain rate of 10^{-5}/sec is almost the same as the response rate of 0.1/sec. When the fiber volume is increased to 0.55, the response of the composites with square fibers at the strain rate of 10^{-5}/sec is even higher than the response of the composites with circular fibers at the strain rate of 0.1/sec.

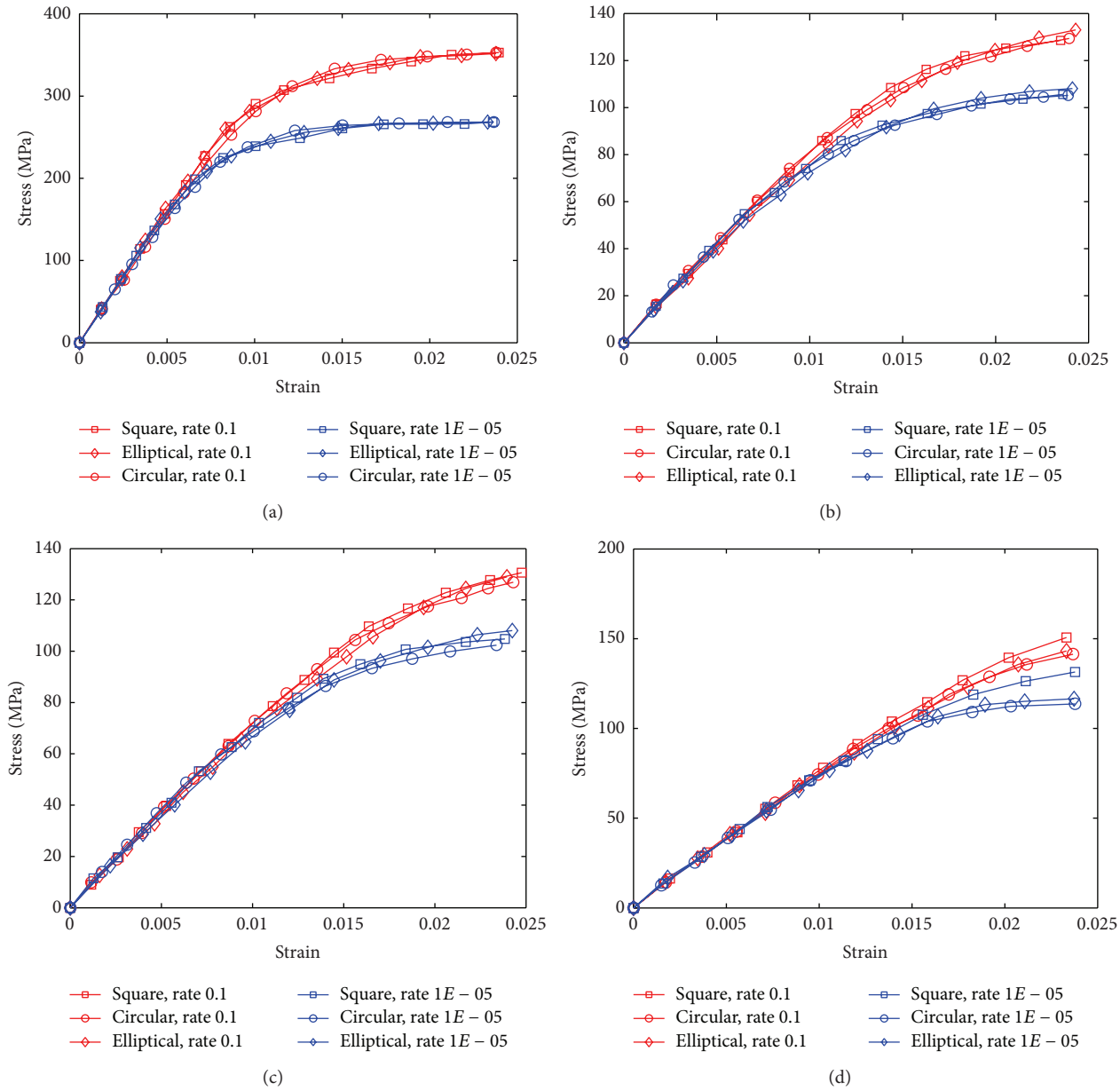


FIGURE 8: Off-axis responses of AS4/PEEK laminate ($v_f = 0.45$) at strain rate of 0.1/sec and 10^{-5}/sec: (a) 15°, (b) 45°, (c) 60°, and (d) 75°.

In Figure 10(d), the square fiber provides a 33% increase in the transverse flow stress of the composite relative to that of the elliptical and circular fibers when the strain rate is 10^{-5}/sec, while the square fiber provides a 15% increase when the strain rate is 0.1/sec.

5. Conclusions

A viscoplastic constitutive model has been employed in the micromechanical method based on generalized model of cells to analyze the inelastic, rate dependent stress-strain response of fiber-reinforced polymer matrix composites with three different microstructures at different fiber off-axis angles condition. The acceptable agreement between the model predictions and experimental results shows that the proposed model can well predict the behaviors of AS4/PEEK composite. At the same time, from the predicted results, the following conclusions are obtained.

(1) The AS4/PEEK composite is a kind of rate dependent material. When the strain rate changes from 10^{-5}/sec to 0.1/sec, the composites provide an effective increase in the flow stress while the elastic behavior almost remain unchanged.

(2) The effects of fiber cross-sectional shape on the behavior of AS4/PEEK composite are related to the fiber volume fraction and fiber off-axis orientation. When the fiber volume fraction is smaller than 0.15, it can be

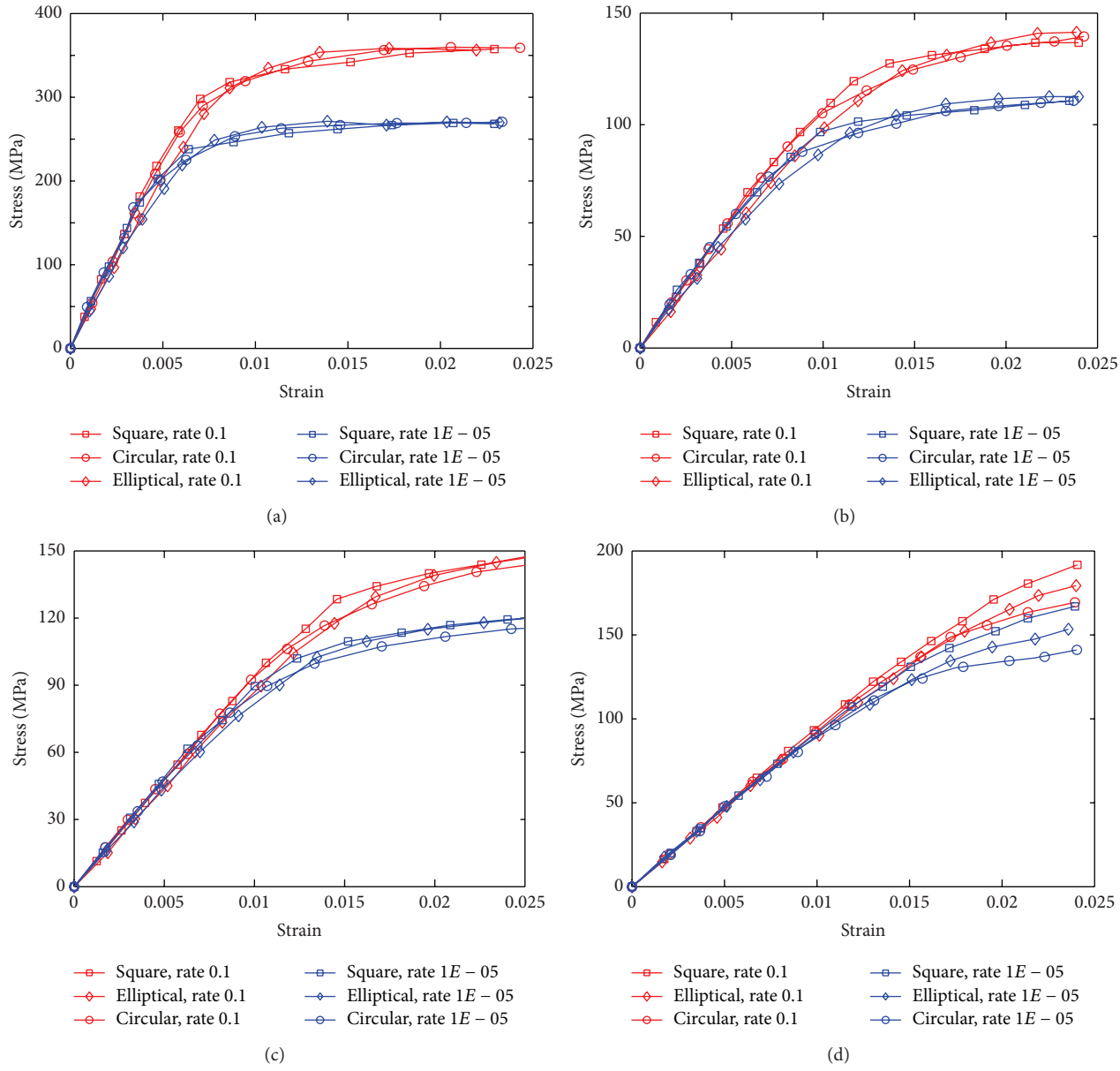


Figure 9: Off-axis responses of AS4/PEEK laminate ($v_f = 0.55$) at strain rate of 0.1/sec and 10^{-5}/sec: (a) 15°, (b) 45°, (c) 60°, and (d) 75°.

seen that the composites response is hardly affected by the fiber cross-section shape; with the increasing of fiber volume fraction and fiber off-axis orientation, the effects of fiber cross-sectional shape become more obvious. Among the three kinds of fiber shapes, the stiffest response is obtained for the composites with the square fibers and the most compliant response for the composites with the circular fibers.

(3) The increasing of fiber volume fraction can improve the stiffness of AS4/PEEK composite. However, for the elliptical fiber, the maximum allowable fiber volume fraction is 0.59 in the case of fibers with an aspect ratio of 4/3, so it should be noted that the elliptical fiber may not be chosen when the fiber volume fraction needed is big.

(4) The influence of fiber off-axis orientation on the stress-strain curves of AS4/PEEK composite is very large. The response of composites decreases obviously when the off-axis orientation changes from 15° to 45° and then increases from 60° to 90°. So when the composites have been chosen to bear the load, the fiber off-axis orientation should be paid attention to.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

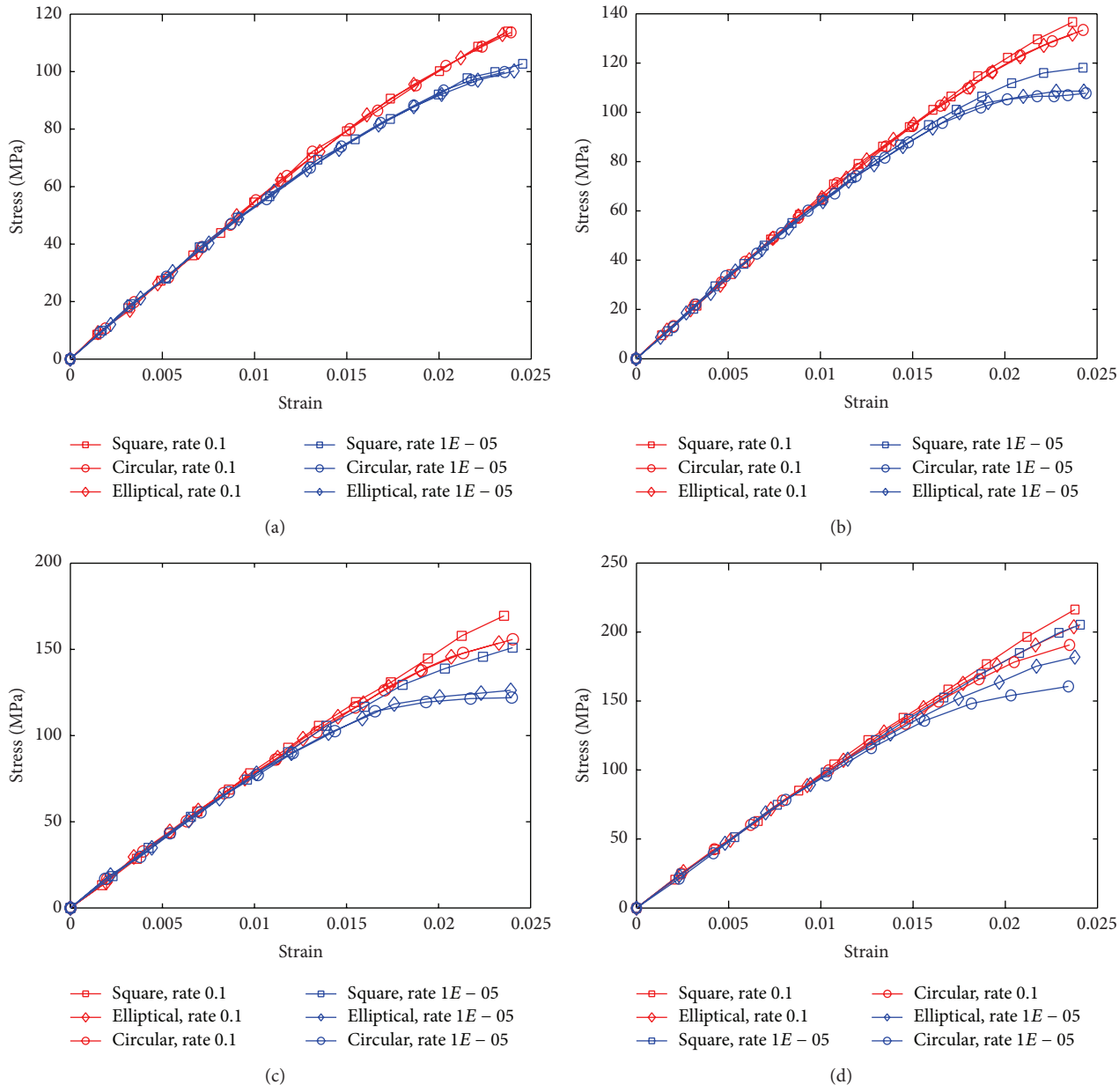


FIGURE 10: Stress-strain response of AS4/PEEK [90°] laminate at strain rate of 0.1/sec and 10^{-5}/sec: (a) $v_f = 0.15$, (b) $v_f = 0.30$, (c) $v_f = 0.45$, and (d) $v_f = 0.55$.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 51175401 and 51335006), the Research Fund for the Doctoral Program of Higher Education of China (no. 20120201110028), and the Program for Changjiang Scholars and Innovative Research Team in University.

References

[1] D. Motamedi, A. Milani, M. Komeili, M. Bureau, F. Thibault, and D. Trudel-Boucher, "A stochastic XFEM model to study delamination in PPS/glass UD composites: effect of uncertain fracture properties," *Applied Composite Materials*, vol. 21, no. 2, pp. 341–358, 2014.

[2] P. W. R. Beaumont, "Targeting problems of composite failure," *Key Engineering Materials*, vol. 417-418, pp. 37–40, 2010.

[3] Y. Zhou and Z. Huang, "A bridging model prediction of the ultimate strength of composite laminates subjected to triaxial loads," *Journal of Composite Materials*, vol. 46, no. 19-20, pp. 2343–2378, 2012.

[4] A. Riccio, A. Raimondo, and F. Scaramuzzino, "A study on skin delaminations growth in stiffened composite panels by a novel numerical approach," *Applied Composite Materials*, vol. 20, no. 4, pp. 465–488, 2013.

[5] R. K. Goldberg, G. D. Roberts, and A. Gilat, "Incorporation of mean stress effects into the micromechanical analysis of the high strain rate response of polymer matrix composites," *Composites B: Engineering*, vol. 34, no. 2, pp. 151–165, 2003.

[6] G. M. Pearce, A. F. Johnson, R. S. Thomson, and D. W. Kelly, "Experimental investigation of dynamically loaded bolted joints in carbon fibre composite structures," *Applied Composite Materials*, vol. 17, no. 3, pp. 271–291, 2010.

[7] C. A. Weeks, *Nonlinear rate dependent response of thick-section composite laminates [Ph.D. thesis]*, Purdue University, 1995.

[8] S. V. Thiruppukuzhi and C. T. Sun, "Models for the strain-rate-dependent behavior of polymer composites," *Composites Science and Technology*, vol. 61, no. 1, pp. 1–12, 2001.

[9] A. Haque and M. Ali, "High strain rate responses and failure analysis in polymer matrix composites—an experimental and finite element study," *Journal of Composite Materials*, vol. 39, no. 5, pp. 423–450, 2005.

[10] M. M. Shokrieh and M. J. Omidi, "Tension behavior of unidirectional glass/epoxy composites under different strain rates," *Composite Structures*, vol. 88, no. 4, pp. 595–601, 2009.

[11] J. Ye, Y. Qiu, Z. Zhai, and X. Chen, "Strain rate influence on nonlinear response of polymer matrix composites," *Polymer Composites*, 2014.

[12] Z. Zhai, Z. He, X. Chen, J. Ye, and X. Zhu, "Fiber cross-section shape effect on rate-dependent behavior of polymer matrix composites with FBGs sensors," *Sensors and Materials*, vol. 25, no. 6, pp. 403–410, 2013.

[13] C. A. Weeks and C. T. Sun, "Modeling non-linear rate-dependent behavior in fiber-reinforced composites," *Composites Science and Technology*, vol. 58, no. 3-4, pp. 603–611, 1998.

[14] S. V. Thiruppukuzhi and C. T. Sun, "Testing and modeling high strain rate behavior of polymeric composites," *Composites B: Engineering*, vol. 29, no. 5, pp. 535–546, 1998.

[15] H. D. Espinosa, H. Lu, P. D. Zavattieri, and S. Dwivedi, "A 3-D finite deformation anisotropic visco-plasticity model for fiber composites," *Journal of Composite Materials*, vol. 35, no. 5, pp. 369–410, 2001.

[16] E. Kontou and A. Kallimanis, "Thermo-visco-plastic behaviour of fibre-reinforced polymer composites," *Composites Science and Technology*, vol. 66, no. 11-12, pp. 1588–1596, 2006.

[17] W. Hufenbach, A. Hornig, B. Zhou, A. Langkamp, and M. Gude, "Determination of strain rate dependent through-thickness tensile properties of textile reinforced thermoplastic composites using L-shaped beam specimens," *Composites Science and Technology*, vol. 71, no. 8, pp. 1110–1116, 2011.

[18] L. Raimondo, L. Iannucci, P. Robinson, and P. T. Curtis, "Modelling of strain rate effects on matrix dominated elastic and failure properties of unidirectional fibre-reinforced polymer-matrix composites," *Composites Science and Technology*, vol. 72, no. 7, pp. 819–827, 2012.

[19] L. Xing, K. L. Reifsnider, and X. Huang, "Progressive damage modeling for large deformation loading of composite structures," *Composites Science and Technology*, vol. 69, no. 6, pp. 780–784, 2009.

[20] J. Ye, X. Chen, Z. Zhai, B. Li, Y. Duan, and Z. He, "Predicting the elastoplastic response of fiber-reinforced metal matrix composites," *Mechanics of Composite Materials*, vol. 46, no. 4, pp. 405–416, 2010.

[21] D. D. Robertson and S. Mall, "Micromechanical relations for fiber-reinforced composites using the free transverse shear approach," *Journal of Composites Technology and Research*, vol. 15, no. 3, pp. 181–192, 1993.

[22] R. K. Goldberg and D. C. Stouffer, "Strain rate dependent analysis of a polymer matrix composite utilizing a micromechanics approach," *Journal of Composite Materials*, vol. 36, no. 7, pp. 773–793, 2002.

[23] A. Tabiei and S. B. Aminjikarai, "A strain-rate dependent micro-mechanical model with progressive post-failure behavior for predicting impact response of unidirectional composite laminates," *Composite Structures*, vol. 88, no. 1, pp. 65–82, 2009.

[24] Z. Huang, "Simulation of the mechanical properties of fibrous composites by the bridging micromechanics model," *Composites A: Applied Science and Manufacturing*, vol. 32, no. 2, pp. 143–172, 2001.

[25] Z.-M. Huang, "Inelastic and failure analysis of Laminate structures by ABAQUS incorporated with a general constitutive relationship," *Journal of Reinforced Plastics and Composites*, vol. 26, no. 11, pp. 1135–1181, 2007.

[26] M. Paley and J. Aboudi, "Micromechanical analysis of composites by the generalized cells model," *Mechanics of Materials*, vol. 14, no. 2, pp. 127–139, 1992.

[27] S. Ogihara, S. Kobayashi, and K. L. Reifsnider, "Characterization of nonlinear behavior of carbon/epoxy unidirectional and angle-ply laminates," *Advanced Composite Materials*, vol. 11, no. 3, pp. 239–254, 2003.

[28] J. Tsai and K. Chen, "Characterizing nonlinear rate-dependent behaviors of graphite/epoxy composites using a micromechanical approach," *Journal of Composite Materials*, vol. 41, no. 10, pp. 1253–1273, 2007.

[29] A. Gilat, R. K. Goldberg, and G. D. Roberts, "Strain rate sensitivity of epoxy resin in tensile and shear loading," *Journal of Aerospace Engineering*, vol. 20, no. 2, pp. 75–89, 2007.

[30] M.-J. Pindera and B. A. Bednarcyk, "An efficient implementation of the generalized method of cells for unidirectional, multiphased composites with complex microstructures," *Composites B: Engineering*, vol. 30, no. 1, pp. 87–105, 1999.

[31] D. C. Stouffer and L. T. Dame, *Inelastic Deformation of Metals: Models, Mechanical Properties, and Metallurgy*, John Wiley & Sons, New York, NY, USA, 1996.

[32] R. K. Goldberg, G. D. Roberts, and A. Gilat, "Implementation of an associative flow rule including hydrostatic stress effects into the high strain rate deformation analysis of polymer matrix composites," *Journal of Aerospace Engineering*, vol. 18, no. 1, pp. 18–27, 2005.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 831075, 6 pages
http://dx.doi.org/10.1155/2014/831075

Research Article

Methods of Fault Diagnosis in Fiber Optic Current Transducer Based on Allan Variance

Lihui Wang,[1,2] Gang Chen,[3] Jianfei Ji,[3] Jian Sun,[3] Jiabin Qian,[1] and Xixiang Liu[1]

[1] *School of Instrument Science and Engineering, Southeast University, Key Laboratory of Micro-Inertial Instrument and Advanced Navigation Technology, Ministry of Education, Nanjing 210096, China*
[2] *State Key Laboratory of Transient Optics and Technology, Xian Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xian 710119, China*
[3] *Jiangsu Electrical Power Company Research Institute, Nanjing 211103, China*

Correspondence should be addressed to Lihui Wang; wlhseu@163.com

Received 24 April 2014; Accepted 31 May 2014; Published 22 June 2014

Academic Editor: Ruqiang Yan

To ensure low failure and high reliability of fiber optic current transducers (FOCTs), it is urgent to study methods of condition monitoring and fault diagnosis in FOCT. Faults in FOCT have statistical characteristics. With the analyzing of time domain and frequency domain features in fiber optic current transformers' measurement data, we establish correspondence between the physical characteristics of key components in transformer and data features and then build diagnostic analysis model based on Allan variance. According to the Allan variance calculation results, we can diagnose fiber optic current transformer's health state and realize faults location. Experiment results show that diagnostic methods based on Allan variance are accurate and effective to identify fault features.

1. Introduction

Fiber optic current transducers (FOCTs) are achieving increased acceptance and application in high voltage substations due to their superior accuracy, bandwidth, dynamic range, and inherent isolation. FOCTs are influenced by various factors such as electricity, heat, machinery, and environment in operation [1, 2], so their performance degrades gradually, which eventually leads to a fault. FOCTs are the basic components of the power system. Once they fail, this will cause local even wide area blackout, resulting in huge economic loss and social impact. At present, domestic and foreign maintenances on power equipment mostly in the regular offline maintenance state, which not only affects the normal work of the power grid but also causes its relatively low efficiency [3]. Thus, it is urgent to research on condition monitoring and fault diagnosis of FOCT.

According to the information about measured values by condition monitoring and their processing results, fault diagnosis of FOCT reasons, judges, finds out fault's types, location, and severity, and then puts forward proposals on equipment repair processing. Condition monitoring is the collection process of characteristic quantity [4]. It records, classifies, and evaluates the running state of the equipment. And it also provides decisions for equipment maintenance and repair. Fault diagnosis is the analysis and judgment process after characteristic quantity's collection, locating the occurred fault and judging the degree of fault development based on fault tag. Due to imperfect characteristics of the optical components in FOCT and environmental interference, there is a series of regular statistical noise in the output signal of FOCT. Random noises of FOCT come from light source coherent noises, light source intensity noises, photodetector shot noises, thermal noises in electronic devices, environmental noises, and time-varying noises because of device aging and other factors. Despite the large number of noise sources, output data of FOCT reflect that the characteristics mainly include angle random walk, bias instability, rate random walk, rate ramp, quantization noise, and sinusoidal noise.

As a complement to the frequency domain analysis, Allan variance is a time domain analysis technique originally

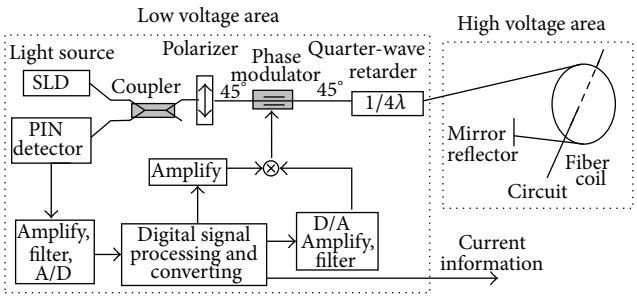


Figure 1: Structure of reflective FOCT with reflection mirror.

developed to study the frequency stability of oscillators [5–7]. It can be used to determine the character of the underlying random processes that give rise to the data noise. As such, it helps to identify the source of a given noise term present in the data, whether it is inherently in the instrument or in the absence of any plausible mechanism within the instrument, its origin should be sought in the test setup.

In this paper, we analyze the principle of Allan variance, evaluate types of random noise error, identify the statistical properties of various types of random errors, and locate noise error source. In the application of condition monitoring and fault diagnosis, we establish relational models between the noise characteristics and fault source with analysis of noise characteristics in time and frequency domain, respectively, according to the characteristics and trend of relevant parameters in FOCT. The model can monitor the health status of FOCT, judge fault position in fault state, provide the alarm when necessary, and finally provide the basis for further steps.

2. Principle of Fault Diagnosis in FOCT

2.1. Introduction of FOCT. FOCT is a kind of optical sensor based on Faraday magneto optical effects and optical interference theory [8, 9]. It uses closed loop feedback system to measure the change of light intensity caused by optical nonreciprocal phase difference in real time, in order to acquire the information of measured current proportional to nonreciprocal phase. FOCT is composed of a wide spectrum light source, a polarization maintaining fiber, a polarizer, an electrooptical modulator, the signal processing unit, and so forth. There are multiple signal transmission and processing links, such as photoelectric conversion and electrooptical conversion. Output data of transducer have the random noise characteristics of FOCT.

According to the structure of fiber optic sensor head, FOCTs are divided into reflective FOCT and Sagnac FOCT [2, 9]. The structure of FOCT with reflection mirror is shown in Figure 1; the structure of FOCT with fiber loop is shown in Figure 2.

Propagation process of light waves in reflective FOCT can be described as follows. The waves generated by the light source are polarized to linear polarized waves by polarizer after passing through the coupler. The linear polarized waves enter the polarization-maintaining fiber at 45°, transferring evenly into the polarization axis X fiber (fast axis) and Y-axis

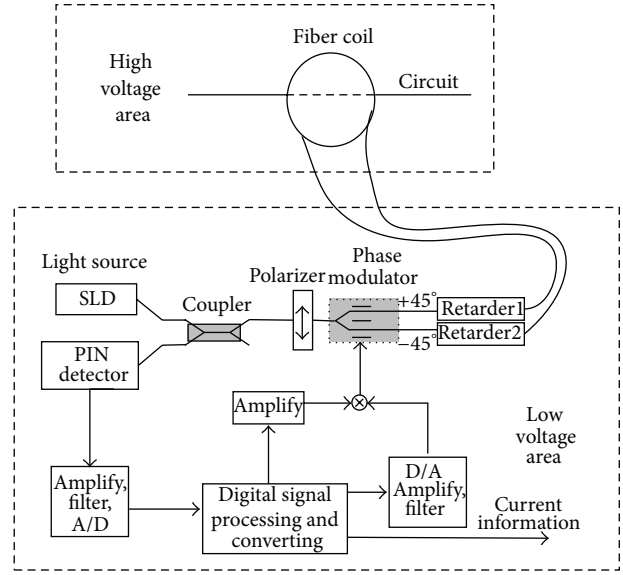


Figure 2: Structure of Sagnac FOCT with fiber loop.

(slow axis). Then, after passing through the 1/4λ wave plate at 45°, the two beams of orthogonal mode waves are converted into left and right circular polarized waves, respectively, which enter the sensor fiber after that. In the sensor fiber, the two circular polarized waves transfer in different speeds due to Faraday field effect caused by current. After mirror reflected by sensor fiber's end face, they exchange their polarization modes (namely, the light of left-handed rotation changes into the light of right-handed rotation, and the light of right-handed rotation changes into the light of left-handed rotation.). They pass through the sensor fiber again and interact with the magnetic field generated by current again, doubling the phase. After passing through the 1/4λ wave plate, the two beams of light revert to linear polarized waves. While they entered along the X-axis and Y-axis, they emit out wave plate along the Y-axis and X-axis of polarization-maintaining fiber, interfering in the polarizer. In the transmission process of the waves, the two beams of interfered waves pass through the X-axis and Y-axis of polarization-maintaining fiber as well as left-handed and right-handed rotation mode of sensor fiber, only different in time. Thus, the waves which return to detector only carry the nonreciprocal phase difference caused by Faraday effect.

Propagation process of light waves in Sagnac FOCT can be described as follows. The waves generated by the light source enter the integrated optical chip through the coupler and are polarized to linear polarized waves through the polarizer. The polarized waves are split into two parts at the beam splitter in the integrated optical chip, entering the wave plate, respectively, at plus or minus 45° (clockwise light is 45° and counterclockwise light is 45°). Prior to entering a low birefringence polarization-maintaining fiber optic sensor head in opposite directions, respectively, the polarized waves are converted into left and right circular polarized waves by 1/4λ wave plate. Faraday field effect caused by current rotates the plane of polarization of two circular polarized waves.

The waves are converted back into linear polarized waves when passing through another $1/4\lambda$ wave plate and are brought back to interference. Rotation angles of polarization plane in two beams of interfered light are equal and opposite, then, interference phase in Sagnac FOCT is twice the Faraday phase shift. According to modulation technology of digital closed-loop signals, the information about the magnitude of current in measured power line can be acquired by detecting the phase difference of output lights.

2.2. Fault Features in FOCT. Although Sagnac FOCT and reflection FOCT are different in structure, they have the similar light waves sensing principle and signal processing steps. That means they all build digital closed-loop feedback system using high speed signal processing unit and electrooptic phase modulator in order to measure the information of nonreciprocal phase caused by Faraday magnetic field effect in real time and finally acquire the information of external current. Influenced by environmental factors and inherent factors, FOCTs reflect the same noise characteristics, which are mainly divided as the following categories [5, 10].

Bias instability (BI) noises reflect the bias low-frequency fluctuation of FOCT. BI noises originate from discharge assembly in FOCTs, plasma discharge, circuit noises, environmental noises, and many other components which can generate random flashing. It is useful to inhibit BI by reliability design of FOCTs and taking corresponding filtering method. Angle random walk (ARW) noises show the ultimate precision of FOCTs and are an important indicator to measure the IFOG noise level. Photon shot noises of photoelectric detector (PIN) in FOCT result in the uncertainty of Faraday phase shift measurement, which cause a limit of current measurement. Shot noises also cause current random fluctuation of current-voltage feedback impedance in PIN preamplifier, resulting in pseudo-Faraday phase shift. The phase shift influences IFOG minimum bias stability and decides FOCT's precision. ARW noises are the result of integrated broadband rate power spectral density, originating mainly from photodetector shot noises, amplifier noises, electronic device thermal noises, and some high frequency noises whose relevant time is shorter than sampling time. High frequency noises whose relevant time is shorter than sampling time can be eliminated. It is also efficient to inhibit ARW noises by using high-qualified light source and photoelectric detector and improving the stability of environmental temperature. Rate ramp (RR) is in essence a definite error, rather than a random error. The strength of light source in FOCT changes monotonously and very slowly and lasts for a long time, which causes RR noises. It is useful to reduce RR noises by ensuring the long-term stability of optoelectronic devices and working environment of FOCTs and determining the error compensation in the method of establishing the mathematical model. Rate random walk (RRW) noises reflect index correlated noises of long correlation time in limit condition. RRW noises are the integral result of phase value power spectral density, associated with long term effects of resonator. They are generated after white noises pass the integrator. RRW can be inhibited by reducing the aging effect of crystal oscillator. Sinusoidal noise (SN) is a kind of systematic error whose power spectral density is presented by several different frequencies. High frequency noises are generated by laser plasma oscillations in the discharge process; low frequency noises are caused by environmental periodic change. When the sinusoidal noises have sinusoidal waveform with multipeak, it is easier to show SN by the plot of power spectral density. Quantization noises (QN) reflect the minimum resolution of current's information of FOCTs. Sampling values of interference signals in FOCT are converted to digital quantities by A/D and are sent into signal processor. In the measurement of time interval, measurement phase induced by current electromagnetic field is not integer times of quantified step size, while the amplitude of signals gets quantified over time, which causes quantization error. In the application environment with requirement of high sampling rate, large QN is caused, which can be reduced by improving the accuracy of acquisition system and shortening the initial sampling time.

2.3. Allan Variance Algorithm and Fault Diagnosis. Supposing that the current data with τ_0 sampling period in FOCT is N-dimensional, we can obtain data collection $\{I(0), I(\tau_0), I(2_{\tau 0}), \ldots, I(N \cdot \tau_0)\}$. Allan variance is a time domain analysis technique based on function of time length. It can be used to calculate frequency data stability of FOCT in the time domain, analyze the random noise characteristic of FOCT, and isolate and identify the random error model and its parameter using the slope of the curve model in log-log plot [5, 6]. The Allan variance is defined as follows:

$$\sigma^2(\tau) = \frac{1}{2}\left\langle \left(\bar{I}_{k+m}(\tau) - \bar{I}_k(\tau)\right)^2 \right\rangle, \tag{1}$$

where $\langle\ \rangle$ is the ensemble average. $\bar{I}_{k+m}(\tau) = (w_{k+2m} - w_{k+m})/\tau$ and $\bar{I}_k(\tau) = (w_{k+m} - w_k)/\tau$ are average current. $\tau = m \cdot \tau_0$ is correlation time. τ_0 is the minimum sampling period. w_k is the value of current for the moment of $k\tau_0$. Consider

$$\begin{aligned}\sigma^2(\tau) &= \frac{1}{2}\left\langle \left(\bar{I}_{k+m}(\tau) - \bar{I}_k(\tau)\right)^2 \right\rangle \\ &= \frac{1}{2\tau^2}\left\langle \left(w_{k+2m} + w_k - 2w_{k+m}\right)^2 \right\rangle.\end{aligned} \tag{2}$$

The Allan variance can be expressed as follows:

$$\sigma^2(\tau) = \frac{1}{2\tau^2(N-2m)} \sum_{k=1}^{N-2m} \left(w_{k+2m} + w_k - 2w_{k+m}\right)^2. \tag{3}$$

The Allan variance obtained by performing the prescribed operations is related to the PSD of the noise terms in the original data set. The relationship between Allan variance and the two-sided PSD $S_E(f)$ is given by the following equation:

$$\sigma^2(\tau) = 4\int_0^\infty S_E(f) \cdot F(f)\, df = 4\int_0^\infty S_E(f) \frac{\sin^4(\pi f t)}{(\pi f t)^2} df. \tag{4}$$

Equation (4) indicates that Allan variance is proportional to the total energy of random noises when the random noises pass through the filter of function $F(f)$. Filter of band pass depends on the correlation time τ. By adjusting it (namely, adjusting band filter), different types of stochastic processes can be detected. The credibility of Allan variance estimation improves with the increasing numbers of independent sets.

Equation (4) is the key result that will be used throughout to characterize the rate noise PSD from the Allan variance calculations. Its physical interpretation is that the Allan variance is proportional to the total noise power of the FOCT current output when passed through a filter with the transfer function of $F(f)$. This particular transfer function is the result of the method used to create and operate on the clusters. It is seen from (4) that the filter of band pass depends on τ. This suggests that different types of random processes can be examined by adjusting the filter of band pass, namely, by varying τ. Thus the Allan variance provides a means of identifying and quantifying various noise terms that exist in the data. It is normally plotted as the square root of the Allan variance versus $\tau[\sigma(\tau)]$, on a log-log plot. The following paragraphs show the application of (4) to a number of noise terms that are either known to exist in the FOCT or otherwise influence its data. Power spectral density of random noise in FOCT has the following features.

For angle random walk, the associated rate PSD is represented by the following equation:

$$S_E(f) = N^2, \tag{5}$$

where N is the angle random walk coefficient.

The following equation is obtained by performing integration:

$$\sigma^2(\tau) = \frac{N^2}{\tau}. \tag{6}$$

For bias instability, the associated rate PSD is represented by the following equation:

$$\begin{aligned} S_E(f) &= \left(\frac{B^2}{2\pi}\right)\frac{1}{f}, \quad f \le f_0, \\ S_E(f) &= 0, \quad f > f_0, \end{aligned} \tag{7}$$

where B is the bias instability coefficient and f_0 is the cut-off frequency.

For rate random walk, the associated rate PSD is represented by the following equation:

$$S_E(f) = \left(\frac{K}{2\pi}\right)^2 \frac{1}{f^2}, \tag{8}$$

where K is the rate random walk coefficient.

For quantization noise, the associated rate PSD is represented by the following equation:

$$S_E(f) = \tau Q^2, \tag{9}$$

where Q is the quantization noise coefficient.

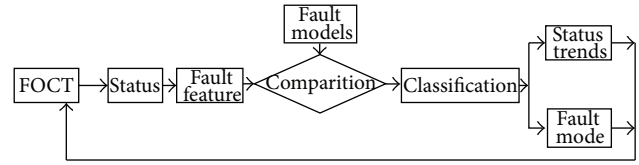


Figure 3: Structure of fault diagnosis system in FOCT.

Different error terms of random noises in FOCT appear in different correlation time domain. It is different in the power spectral density and the function relationship with correlation time τ of different noises. Assuming that the noise error term is independent in the statistical sense, integrated Allan variance $\sigma^2_{\text{total}}(\tau)$ is given by the following equation:

$$\sigma^2_{\text{total}}(\tau) = \sum_{n=-2}^{2} A_n \tau^n, \tag{10}$$

where A_n ($n = -2, -1, 0, 1, 2$) correspond to quantization noise, angle random walk, bias stability, rate random walk and rate ramp, and many other coefficients of fitted polynomial related to noise, respectively. Consider

$$\begin{aligned} \sigma^2_{\text{total}}(\tau) = \sum_{n=-2}^{2} A_n \tau^n &= \frac{3Q^2}{\tau^2} + \frac{N^2}{\tau} \\ &+ \frac{2\ln 2}{\pi} B^2 + \frac{K^2}{3}\tau + \frac{R^2}{2}\tau^2. \end{aligned} \tag{11}$$

In (11), Q is quantization noise coefficient; N is angle random walk coefficient, B is bias instability coefficient; K is rate random walk coefficient; R is rate ramp coefficient. Each type of noise errors corresponds to different slope in the Allan variance correlation time log-log plot. The slope of quantization noise, angle random walk, bias stability, rate random walk, and rate ramp is −1, −0.5, 0, 0.5 and 1, respectively [5, 6].

Allan variance analysis method and the modeling technology can effectively separate and identify several main types of random noises which influence FOCT's precision. By analyzing the curve, the corresponding noise error values can be estimated, making a comprehensive evaluation of overall performance of FOCT. In addition, error source of FOCT's noise error can be located by analyzing different types of noise values. It is not only effective to improve the performance of FOCT, but also convenient to identify and locate fault in FOCT.

Combining FOCT state parameters and available information, we can build an intelligent diagnosis system based on some algorithms, to determine the status of the device with theoretical derivation and realize fault detection and diagnosis [11–13]. The structure of intelligent diagnosis system is shown in Figure 3. Figure 4 shows fault diagnosis process in FOCT. We can obtain feature vectors in time domain and frequency domain with signal processing model and then detect the failure source by using the mapping relationship between feature vectors and failure source.

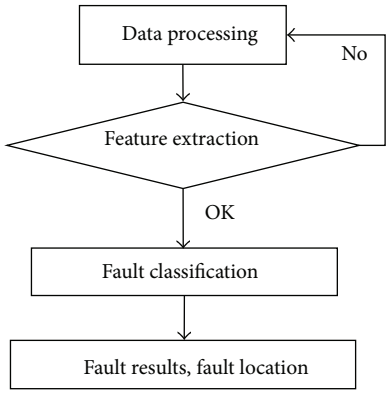


Figure 4: Flow chart of fault diagnosis in FOCT.

Figure 5: Experimental environment of current transducer test system.

3. Experiments

To verify the performance and fault diagnosis methods of fiber optic current transducer, we built a current transformer test system, as shown in Figure 5. Figure 6 shows the principle of current transformer test system. FOCT, standard current transducer (CT), and current generator are stringed in the same current loop; the current generator applies primary current to the two sets of CT, while different error factors are added to FOCT. The function of transducer calibration part is to process data in FOCT and reference standard CT and to calculate the measurement deviation. The function of fault diagnosis part is to monitor FOCT's operation status and to diagnose FOCT's fault by using Allan variance methods.

We test fault characteristics of several fiber optic current transformers, respectively. Figure 7 shows sample of current measurements data in one FOCT. Allan variance log-log curve expresses the random error in fiber current transformer accurately, including angle random walk, bias instability, rate random walk, rate ramp, quantization noise, and sinusoidal noise. Theoretically, the slope of quantization noise, angle random walk, bias stability, rate random walk, and rate ramp is −1, −0.5, 0, 0.5, and 1, respectively. Based on characteristics of random error, we can evaluate health state of FOCT; moreover, error source in FOCT is located. Figure 8 shows Allan variance analysis curve of the first FOCT. We can find that quantization noise with rate ramp of −1, angle random walk with rate ramp of −1/2, and bias instability noise with

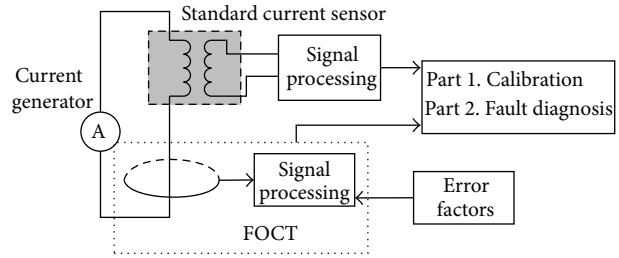


Figure 6: Principle block of current transducer test system.

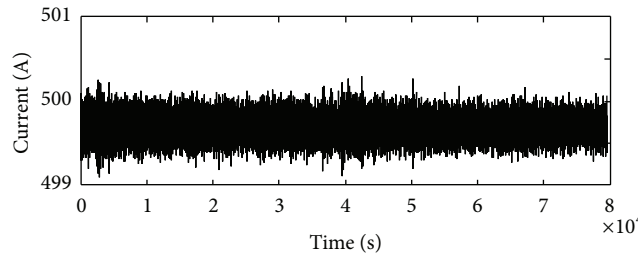


Figure 7: Current measurements data in FOCT.

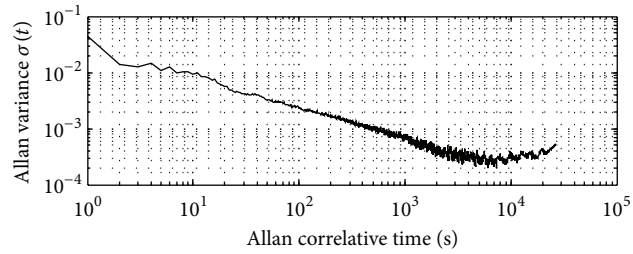


Figure 8: Allan variance curve of first FOCT.

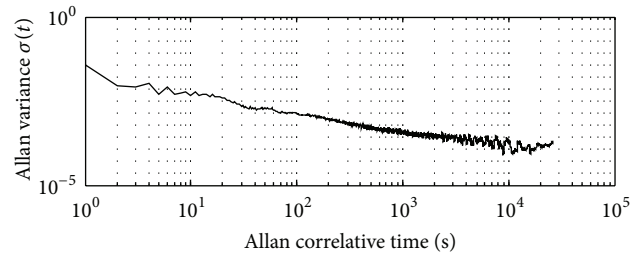


Figure 9: Allan variance curve of second FOCT.

rate ramp of 0 are the main errors in FOCT; moreover, we can locate noise source, mainly from photon shot noise in detector, amplifier noise, thermal noise electronics, and high frequency noise. Minimum resolution performance of this FOCT is affected. Figure 9 shows Allan variance analysis curve of the second FOCT. We can find that angle random walk with rate ramp of −1/2 and sinusoidal noise are the main errors in FOCT, and high-frequency noise dominated system noise is the main noise source.

4. Conclusions

Allan variance can be used to analyze the time domain characteristics of the underlying random processes in FOCT. It helps to identify the source of a given noise term present in the data, whether it is inherently in FOCT or in the absence of any plausible mechanism within FOCT. In this paper, we verified the validity of the Allan variance methods in soft fault diagnosis through theoretical analysis and fault simulation experiments. Allan variance is suitable for evaluation and diagnosis of soft fault in FOCT, which can avoid potential failures, such as performance fault induced by temperature, light source, and other factors. For abrupt-changing fault, we can compare and judge the fault characteristics with feature model by using wavelet transform methods and then decide whether the fault is from FOCT or grid failure.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The project is supported by the following funds. Fundamental Research Funds for the Central Universities (2242013R30016), Natural Science Foundation of Jiangsu Province (BK2012326, BK20130099), National Natural Science Foundation of China (61203192), Research Fund of China Ship 8 Industry (13J3.8.4), and Foundation of Key Laboratory of Micro-Inertial Instrument and Advanced Navigation Technology, Ministry of Education (201103).

References

[1] K. Bohnert, P. Gabus, J. Kostovic, and H. Brändle, "Optical fiber sensors for the electric power industry," *Optics and Lasers in Engineering*, vol. 43, no. 3–5, pp. 511–526, 2005.

[2] J. Blake, P. Tantaswadi, and R. T. de Carvalho, "In-line sagnac interferometer current sensor," *IEEE Transactions on Power Delivery*, vol. 11, no. 1, pp. 116–121, 1996.

[3] H.-B. Wang, K.-M. Tang, R.-L. Xu, X.-J. Zhu, and X.-J. Li, "Diagnosis of soft fault of electronic transformer in digital substation," *Power System Protection and Control*, vol. 40, no. 24, pp. 53–58, 2012.

[4] S. N. Huang and K. K. Tan, "Fault detection, isolation, and accommodation control in robotic systems," *IEEE Transactions on Automation Science and Engineering*, vol. 5, no. 3, pp. 480–489, 2008.

[5] "IEEE standard specification format guide and test procedure for single-axis laser gyros," Tech. Rep. IEEE Std 647-2006, IEEE Aerospace and Electronic Systems Society.

[6] "IEEE standard specification format guide and test procedure for single-axis interferometric fiber optic gyross," Tech. Rep. IEEE Std 952-1997, Gyro and Accelerometer Panel of the IEEE Aerospace and Electronic Systems Society.

[7] D. W. Allan and J. A. Barnes, "A modified Allan variance with increased oscillator characterization ability," in *Proceedings of the 35th Annual Frequency Control Symposium*, vol. 5, pp. 470–475, 1981.

[8] K. Bohnert, P. Gabus, J. Nehring, and H. Brändle, "Temperature and vibration insensitive fiber-optic current sensor," *Journal of Lightwave Technology*, vol. 20, no. 2, pp. 267–276, 2002.

[9] K. Bohnert, P. Gabus, J. Nehring, H. Brändle, and M. G. Brunzel, "Fiber-optic current sensor for electrowinning of metals," *Journal of Lightwave Technology*, vol. 25, no. 11, pp. 3602–3609, 2007.

[10] N. Zhang and X. Li, "Research on theoretical improvement of dynamic Allan variance and its application," *Acta Optica Sinica*, vol. 31, no. 11, Article ID 1106003, 2011.

[11] K. Bouibed, A. Aitouche, and M. Bayart, "Sensor and actuator fault detection and isolation using two model based approaches: application to an autonomous electric vehicle," in *Proceedings of the 18th Mediterranean Conference on Control & Automation (MED '10)*, pp. 1290–1295, Marrakech, Morocco, June 2010.

[12] K. Bouibed, A. Aitouche, and M. Bayart, "Nonlinear parity space applied to an autonomous vehicle," *Journal of Energy and Power Engineering*, vol. 3, no. 12, pp. 10–18, 2009.

[13] A.-J. Khalid, J. Wang, and M. Nurudeen, "A new fault classification model for prognosis and diagnosis in CNC machine," in *Proceedings of the 25th Chinese Control and Decision Conference (CCDC '13)*, pp. 3538–3543, May 2013.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 696051, 11 pages
http://dx.doi.org/10.1155/2014/696051

Research Article

Sparse Representation of Transients Based on Wavelet Basis and Majorization-Minimization Algorithm for Machinery Fault Diagnosis

Wei Fan,[1] Gaigai Cai,[1,2] Weiguo Huang,[1] Li Shang,[3] and Zhongkui Zhu[1,2]

[1] *School of Urban Rail Transportation, Soochow University, Suzhou 215137, China*
[2] *State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China*
[3] *College of Electronic Information, Suzhou Vocational University, Suzhou 215104, China*

Correspondence should be addressed to Zhongkui Zhu; zkzhu@ustc.edu

Received 5 April 2014; Accepted 21 May 2014; Published 19 June 2014

Academic Editor: Ruqiang Yan

Vibration signals captured from faulty mechanical components are often associated with transients which are significant for machinery fault diagnosis. However, the existence of strong background noise makes the detection of transients a basis pursuit denoising (BPD) problem, which is hard to be solved in explicit form. With sparse representation theory, this paper proposes a novel method for machinery fault diagnosis by combining the wavelet basis and majorization-minimization (MM) algorithm. This method converts transients hidden in the noisy signal into sparse coefficients; thus the transients can be detected sparsely. Simulated study concerning cyclic transient signals with different signal-to-noise ratio (SNR) shows that the effectiveness of this method. The comparison in the simulated study shows that the proposed method outperforms the method based on split augmented Lagrangian shrinkage algorithm (SALSA) in convergence and detection effect. Application in defective gearbox fault diagnosis shows the fault feature of gearbox can be sparsely and effectively detected. A further comparison between this method and the method based on SALSA shows the superiority of the proposed method in machinery fault diagnosis.

1. Introduction

Detection of transients has always been an important task in image and signal processing, and it has also been increasingly significant in the field of machinery fault diagnosis in recent years [1–3]. Most of the machinery is operated by means of gears and other rotating parts. Fault in these parts may lead to the failure of the whole machine and even loss of lives. Thus, proper analysis for transient detection from the machinery vibration signals has attracted sustained attention during the past decades [4–8].

Vibration signals derived from the defective gearboxes are generally observed as periodic transient impulses due to the rotating status [9]. Researches have shown that the transients in the vibration signals generated from a machinery component usually correspond to the localized fault or mechanical defect, such as flaking, crack, breakage, and fracture [10]. Therefore, it is very important to extract the useful information, such as the feature of transient impulse and the cycle of multiple transient impulses, by analyzing the transients in the vibration signal [11, 12]. However, the strong background noise in practical vibration signals will corrupt the fault-induced transient impulses. Hence, detection of transients hidden in the noisy vibration signals is of great importance for machinery fault diagnosis.

Different methods have been proposed to extract the fault features from the vibration signals, such as wavelet analysis [13], time-frequency analysis [14, 15] and empirical mode decomposition (EMD) [16], and so forth. Meanwhile, another feature extraction method based on sparse representation is widely used in image processing, which has a close analogy with the effect of those fault feature extraction methods for 1-D signal.

Generally, transients to be detected can be described as a series of sparse coefficients multiplied by a certain wavelet basis as long as the signal being processed has

a sparse representation with respect to the known wavelet basis [17]. Hence these signal processing problems can be viewed as a series of restoration and reconstruction problems, which are classical linear inverse problems. However, most inverse problems are ill-posed [18, 19]; thus, these inverse problems can only be solved satisfactorily by adopting some sort of regularization. Based on sparse representation theory, a standard formulation for coping with these problems called basis pursuit denoising (BPD), consisting of a data fidelity term and a penalty term, has been proposed in [20]. One approach to derive suitable algorithms is based on the majorization-minimization (MM) method from the optimization theory. To apply the MM algorithm, either the data fidelity term (a quadratic function) or the penalty term (usually nonquadratic) can be majorized to handle this class of BPD problems. Typical approach for solving the former term is the so-called split augmented Lagrangian shrinkage algorithm (SALSA), which has been proposed recently to handle the image recovery problems [21]; while, for the latter term, an algorithm associated with an alternative quadratic function employing majorization-minimization also provides the solution to the corresponding inverse problems [22], called MM algorithm based on nonquadratic majorization.

SALSA, a recent algorithm to solve the BPD problem by using the Hessian of the data fidelity term, has better convergence properties than the earlier algorithms, such as iterative shrinkage/thresholding algorithm (ISTA) and fast ISTA (FISTA) [23, 24]. Based on variable splitting, this algorithm obtains an equivalent constrained optimization formulation, which is then addressed with an augmented Lagrangian method, and then, the convergence is guaranteed. The transient feature detection method based on SALSA has been increasingly popular at present. Sparsity-enabled signal decomposition method based on SALSA is used to decompose the vibration signal of faulty gearbox into high-oscillatory component and low-oscillatory component, which is employed by Cai et al. to extract the fault feature of gearbox [25]. Application of the sparse algorithm SALSA is employed by Selesnick to extract the frequency component of a sinusoidal signal [26]. However, some problems still remain. For instance, it is hard to select the optimal basis for signal representation to avoid the loss of useful components and the computation cost is large because two quadratic functions associated with nonconvex regularization functions should be minimized during each iteration of this algorithm.

Instead of utilization of variable splitting, the MM algorithm based on nonquadratic majorization utilizes a sequence of simpler convex optimization problems to replace the original ill-posed inverse problems yet is an effective and widely applicable method [27]. The algorithm has wide applications in image deconvolution or restoration under certain regularization due to its flexibility in the design of the sequence of simpler optimization problems so does in the field of speech denoising [22, 28–30]. However, there are still some remaining issues to be studied when applying the MM algorithm. One of the most important issues is how to select the optimal wavelet basis to satisfy the characteristics of the machinery vibration signals, which is very close to the important issue of SALSA, and this is the vital point to improve the detection performance.

Considering that the wavelet basis of nonquadratic majorization-based MM algorithm in recent image processing is not suitable for feature extraction of gearbox fault vibration signal, and the selection of appropriate wavelet basis is the premise of the detection effect, a new method must be proposed to accommodate to the actual signal. Mass data show that the waveform of Morlet wavelet is in shape similar to signal transients caused by gearbox localized defects [3, 31]; thus, this paper proposes a new technique by synthesizing MM algorithm and Morlet wavelet basis. To implement the sparsity of the results, optimal wavelet basis is selected by applying correlation filtering, which has been widely applied in mechanical vibration signal processing [32]. The key issue to minimize the objective function of the ill-posed inverse problem is solved by nonquadratic majorization-based MM algorithm; thus a meaningful sparse coefficient would be got. Finally, the transients can be detected from the sparse coefficients, which are the key features for machinery fault diagnosis.

The remainder of the paper is organized as follows. In Section 2, the basic theoretical background concerning the proposed method is introduced. Section 3 gives a simulated study and analysis to verify the proposed method. Section 4 applies the presented method to the fault diagnosis of a gearbox in an automobile transmission gearbox, with a comparison to the extraction results of SALSA method. Finally, conclusions are drawn in Section 5.

2. Theory and Method

In transient detection problems, the goal is to detect the transients buried in noise from an observed signal. The transients can be represented as a series of sparse coefficients as long as the signal being analyzed has a sparse representation with respect to a known transformation. In this section, MM algorithm, solving ill-posed inverse problems, is introduced to model the transient detection problem. The following part provides a brief description on main steps of the MM-based transient detection method.

2.1. Mathematical Model of Transient Detection Problem. Considering that the signal sampled from the transducer always has much noise, the observed signal $y(t)$ can be modeled as

$$y(t) = x(t) + n(t), \tag{1}$$

where $y(t)$ is the sampled signal from the transducer, $x(t)$ is the true signal without noise, and $n(t)$ is the noise. The true signal $x(t)$ can be represented as a sparse linear combination of certain atoms. The representation of x can be expressed as $x = Ac$, where c is the vector of representation coefficients and also represents the transients; and the set of columns of A is a wavelet basis or dictionary; to be convenient, we also call A as wavelet basis. With this sparse representation, the estimation model (1) becomes

$$y(t) = Ac + n(t), \tag{2}$$

where A is an $N \times M$ matrix, $x = Ac$ is a length-N vector, and c is a length-M vector, with $M > N$. The more the similarity between the basis A and the signal x is, the sparser the vector c will be. Here assume that AA^* (where $(\cdot)^*$ denotes complex conjugate transpose) is invertible; therefore, the system of (2) has infinitely many solutions. As in many recent publications [21, 22, 26], we adopt the l_1-norm regularizer to handle the ill-posed nature of the problem of inferring c and thus leads to the BPD problem:

$$\arg\min_c F(c) \quad \text{with } F(c) = \frac{1}{2}\|y - Ac\|_2^2 + \lambda\|c\|_1, \tag{3}$$

where $\|c\|_1$ is the l_1-norm of M-length vector c, defined as $\|c\|_1 = \sum_{m=1}^{M} |c(m)|$, $\|c\|_2$ is the l_2-norm of M-length vector c, defined as $\|c\|_2^2 = \sum_{m=1}^{M} |c(m)|^2$, and λ is the regularization parameter.

The function $F(c)$ cannot be easily minimized unless it is quadratic. To minimize $F(c)$, an iterative algorithm must be used. According to the MM algorithm from the optimization theory, instead of minimizing $F(c)$, the MM algorithm solves a sequence of simpler minimization problems:

$$c_{k+1} = \arg\min_c G_k(c), \tag{4}$$

where k is the iteration counter, $k = 1, 2, 3, \ldots$. The MM algorithm asks that each function $G_k(c)$ should be a majorizer (upper bound) of $F(c)$ and that it coincides with $F(c)$ at $c = c_k$. That is

$$\begin{aligned} &\forall c, G_k(c) \geq F(c), \\ &G_k(c_k) = F(c_k). \end{aligned} \tag{5}$$

2.2. Majorization Iterative Algorithm Based on MM Approach. Traditional MM approaches to this problem consider three possible majorization strategies and thus lead to three different classes of algorithms as shown in [16]. As mentioned above, we can solve this problem easier if the function is quadratic.

Note that the data fidelity term $\|y - Ac\|_2^2$, presented in the function $F(c)$, is already a quadratic function. Then we just need to focus our attention on the penalty term $\|c\|_1$.

We mark penalty term $\|c\|_1$ as $\Psi(c)$, term $|c|$ as $\phi(c)$; then $\Psi(c) = \|c\|_1 = \sum_{m=1}^{M} |c(m)| = \sum_{m=1}^{M} \phi(c(m))$. $\phi(c)$ is an absolute value function and thus is nondifferentiable and nonstrictly convex, which makes (3) difficult to solve. According to the MM algorithm as shown in (5), we should find a quadratic function $g(c)$ to majorize $\phi(c)$ of the general form:

$$g(c) = mc^2 + nc + b, \tag{6}$$

where the parameters m, n and b are constants, the majorizer $g(c)$ should be the upper bound for $\phi(c)$ that coincides with $\phi(c)$ at a specified point c_k as shown in Figure 1. For this quadratic majorizer, conditions in (5) are equivalent to

$$\begin{aligned} &g(c_k) = \phi(c_k), \\ &g'(c_k) = \phi'(c_k). \end{aligned} \tag{7}$$

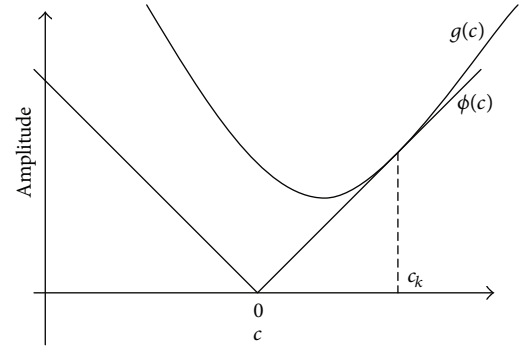


Figure 1: The penalty function and its quadratic majorizer.

Solving for m and b gives $m = (\phi'(c_k)/2c_k) - (n/2c_k)$, $b = \phi(c_k) - (c_k/2)\phi'(c_k) - (n/2)c_k$, thus leading to the majorizer $g(c)$ in (6) given by

$$\begin{aligned} g(c) = &\left(\frac{\phi'(c_k)}{2c_k} - \frac{n}{2c_k}\right)c^2 + nc \\ &+ \left(\phi(c_k) - \frac{c_k}{2}\phi'(c_k) - \frac{n}{2}c_k\right). \end{aligned} \tag{8}$$

Considering a special form of function $g(c)$, we set the unknown parameter $n = 0$; then the parameters m and b become $m = (\phi'(c_k)/2c_k)$, $b = \phi(c_k) - (c_k/2)\phi'(c_k)$, thus leading to the majorizer $g(c)$ in (8) given by

$$g(c) = \frac{\phi'(c_k)}{2c_k}c^2 + \phi(c_k) - \frac{c_k}{2}\phi'(c_k). \tag{9}$$

Considering the concrete form of $\phi(c)$, the function $G_k(c)$ can be written in a matrix format as

$$G_k(c) = \frac{1}{2}\|y - Ac\|_2^2 + \lambda\left(\frac{1}{2}c^*\Lambda_k^{-1}c + \frac{1}{2}\|c_k\|_1\right), \tag{10}$$

where Λ_k denotes the diagonal matrix with vector $|c_k|$ along its diagonal,

$$\Lambda_k = \operatorname{diag}(|c_k|) = \begin{bmatrix} |c_k(0)| & & \\ & \cdots & \\ & & |c_k(M-1)| \end{bmatrix}. \tag{11}$$

Then, the MM update (4) for c_k is

$$c_{k+1} = \arg\min_c \left[\frac{1}{2}\|y - Ac\|_2^2 + \lambda\left(\frac{1}{2}c^*\Lambda_k^{-1}c + \frac{1}{2}\|c_k\|_1\right)\right]. \tag{12}$$

The last term of (12) can be omitted because it does not depend on c; thus a new update equation also called cost function is got as

$$c_{k+1} = \arg\min_c \frac{1}{2}\|y - Ac\|_2^2 + \frac{\lambda}{2}c^*\Lambda_k^{-1}c. \tag{13}$$

Equation (13) is quadratic in c so the solution to this problem can be written explicitly using linear algebra as

$$c_{k+1} = \left(A^*A + \lambda\Lambda_k^{-1}\right)^{-1}A^*y. \tag{14}$$

Although (14) is mathematically valid, there are still some remaining issues with this update. Due to the sparsity of c, components of c will go to zero as the iteration progresses. Thus the first issue is that as the elements of c go to zero, elements of Λ_k^{-1} would go to infinity, which will make the update numerically inaccurate. The second important issue is the selection of the optimal wavelet basis to ensure the sparsity of c when applying MM algorithm to transient detection problem.

The first issue can be avoided, as described in [22], by using the matrix inverse lemma, so the update equation can be written as

$$c_{k+1} = \frac{1}{\lambda}\Lambda_k \left[A^* y - A^* \left(A\Lambda_k A^* + \lambda I\right)^{-1} A\Lambda_k A^* y\right], \quad (15)$$

where λ can be chosen according to the regularization in [30].

The second important issue, the selection of the optimal wavelet basis, can be avoided by employing correlation filtering, which will be described in the next section.

2.3. Selection of Optimal Basis by Correlation Filtering. Morlet wavelet is one of the most popular nonorthogonal wavelets, defined in the time domain as a harmonic wave multiplied by a Gaussian time domain window:

$$\psi(t) = \exp\left(-\frac{\beta^2 t^2}{2}\right)\cos(\pi t). \quad (16)$$

It is a cosine signal that decays exponentially on both the left and the right sides. This feature makes it very similar to an impulse. It has been used for impulse isolation and mechanical fault diagnosis [3, 32–34].

Considering that the waveform of Morlet wavelet is in shape similar to the signal transients caused by gearbox localized defects at a constant speed, we incorporate Morlet wavelet into the above algorithm. The Morlet wavelet is firstly chosen as the basis.

The parametric formulation of Morlet wavelet is

$$\begin{aligned}\psi(f,\zeta,\tau,t) &= \psi_\gamma(t)\\ &= e^{\left(-\zeta/\sqrt{1-\zeta^2}\right)[2\pi f(t-\tau)]^2}\cos\left(2\pi f(t-\tau)\right),\end{aligned} \quad (17)$$

where parameter vector $\gamma = (f,\zeta,\tau)$ determines the wavelet properties, these parameters (f,ζ,τ) are denoted frequency $f \in R^+$, damping ratio $\zeta \in [0,1) \subset R^+$, and time index $\tau \in R$, respectively.

The discrete parameters f, ζ, and τ belong to the subsets of F, Z, and T_C, respectively:

$$\begin{aligned}F &= \{f_1, f_2, \ldots, f_i\} \subset R^+,\\ Z &= \{\zeta_1, \zeta_2, \ldots, \zeta_j\} \subset R^+ \cap [0,1),\\ T_C &= \{\tau_1, \tau_2, \ldots, \tau_k\} \subset R.\end{aligned} \quad (18)$$

With different parameters, dictionary of Morlet wavelet can be constructed as follows:

$$\begin{aligned}\Psi &= \left\{\psi_\gamma(t) : \gamma \in F \times Z \times T_C\right\}\\ &= \left\{\psi(f,\zeta,\tau,t) : f \in F,\ \zeta \in Z,\ \tau \in T_C\right\},\end{aligned} \quad (19)$$

and each item in the dictionary is called an atom.

With the constructed dictionary, correlation filtering is exploited to identify the optimal set of parameters $(\overline{f},\overline{\zeta},\overline{\tau})$, which is the most similar to the original signal [32, 34]. Correlation, which shows the similarity between the basis and the original signal, can be measured by an inner product operation. Then, a correlation function c_γ is defined to quantify the correlation degree between the basis $\psi_\gamma(t)$ and the original signal $x(t)$;

$$c_\gamma = \cos\theta = \frac{\left|\left\langle\psi_\gamma(t), x(t)\right\rangle\right|}{\left\|\psi_\gamma(t)\right\|_2 \|x(t)\|_2}, \quad (20)$$

where θ is the angle between $\psi_\gamma(t)$ and $x(t)$, and the smaller the angle, the nearer the $\psi_\gamma(t)$ and $x(t)$ in space and the more resemble the $\psi_\gamma(t)$ and $x(t)$. Thus, the single similar wavelet function $\psi(t,\overline{f},\overline{\zeta},\tau)$ with optimal parameters $(\overline{f},\overline{\zeta},\tau)$ can be found by maximizing the correlation function c_γ at each time value from the constructed Morlet wavelet dictionary. Peaks of c_γ for a given time value τ can be represented as

$$k_\gamma(\tau) = \max_{f\in F,\zeta\in Z} c_\gamma = c\left(\overline{f},\overline{\zeta},\tau\right), \quad (21)$$

which relates the wavelets with the strongest correlation to the signal, and the time index parameter $\overline{\tau}$ can be found by maximizing the coefficient $k_\gamma(\tau)$; thus the most similar wavelet function is $\psi(t,\overline{f},\overline{\zeta},\overline{\tau})$.

The optimal wavelet basis $\mathrm{Mor}(t,\tau)$ with optimal parameters $\overline{f},\overline{\zeta}$ based on $\psi(t,\overline{f},\overline{\zeta},\overline{\tau})$ replacing the above basis A in (15) can be constructed; thus the second important issue of the MM algorithm has been solved.

2.4. Sparse Representation of Transients Based on Wavelet Basis and MM Algorithm. Motivated by the merits of MM algorithm, a new sparse representation of transients method based on MM algorithm incorporated wavelet basis is proposed in this study. As the signal being analyzed can be represented sparsely on certain wavelet basis, the transients detection problem can be viewed as a basis pursuit denoisng problem; thus the MM algorithm can be employed to handle it.

Considering that the faulty gearbox vibration signal always performs as periodic double-side attenuation functions due to the rotating nature, which is in shape similar to the Morlet wavelet, this paper inserts Morlet wavelet basis $\mathrm{Mor}(t,\tau)$ selected by correlation filtering into the MM algorithm and thus yields the iterative algorithm as follows:

$$\begin{aligned}&c_{k+1}\\ &= \frac{1}{\lambda}\Lambda_k\left[\mathrm{Mor}^* y - \mathrm{Mor}^*\left(\mathrm{Mor}\Lambda_k\mathrm{Mor}^* + \lambda I\right)^{-1}\right.\\ &\qquad\left.\times\ \mathrm{Mor}\Lambda_k\mathrm{Mor}^* y\right].\end{aligned} \quad (22)$$

The updating of (22) can be implemented as follows:

(1) set $k = 0$, choose λ, and initialize c;

(2) set iterative number Nit;

Table 1: Procedures of these two methods: the proposed algorithm and wavelet-assisted SALSA algorithm.

The proposed algorithm	**Wavelet-assisted SALSA** Algorithm
(1) Set k, c_0, λ	(1) Set $k = 0$, v_0, d_0, c_0, λ, μ
(2) **Repeat**	(2) **Repeat**
$\Lambda_k = \mathrm{diag}(\lvert c_k \rvert)$	$c_{k+1} = \arg\min_c \lVert \mathrm{Mor} \cdot c_k - y \rVert_2^2 + \mu \lVert c_k - v_k - d_k \rVert_2^2$
$c_{k+1} = \arg\min_c \left(\frac{1}{2}\right) \lVert y - \mathrm{Mor} \cdot c_k \rVert_2^2 + \left(\frac{\lambda}{2}\right) c_k{}^* \Lambda_k^{-1} c_k$	$v_{k+1} = \arg\min_v \lambda \lVert v_k \rVert_1 + \left(\frac{\mu}{2}\right) \lVert c_{k+1} - v_k - d_k \rVert_2^2$
$k \longleftarrow k + 1$	$d_{k+1} = d_k - (c_{k+1} - v_{k+1})$
until stopping criterion is satisfied.	$k \longleftarrow k + 1$
	until stopping criterion is satisfied.

(3) **repeat**

$$
\begin{aligned}
g &= \frac{1}{\lambda} \mathrm{Mor}^* y \\
\Lambda_k &= \mathrm{diag}(c_k) \\
F &= \lambda I + \mathrm{Mor} \Lambda_k \mathrm{Mor}^* \\
c_{k+1} &= \Lambda_k \left(g - \mathrm{Mor}^* F^{-1} \mathrm{Mor} \Lambda_k g \right) \\
k &\longleftarrow k + 1
\end{aligned}
\tag{23}
$$

until stopping criterion is satisfied ($k >$ Nit).

The matrix Mor has two parameters $\overline{f}, \overline{\zeta}$ as defined above. Then the reconstructed signal $\widehat{x}$ can be expressed as $\widehat{x} = \mathrm{Mor} \cdot \widehat{c}$, where $\widehat{c}$ is the final vector with sparse coefficients.

To illustrate the superiority of the computation cost of the proposed method, the procedures of this method and another sparse algorithm SALSA also based on Morlet wavelet basis are shown in Table 1. As shown in Table 1, wavelet-assisted SALSA algorithm has two convex functions to be minimized, and one of them includes a nonstrictly convex function, which is very hard to minimize. However, there is only one function to be minimized in the proposed method, and the function is quadratic, which makes the proposed method less computation cost.

In summary, the procedure of the proposed transient sparse representation method is illustrated in Figure 2. After the accomplishment of vibration signal measurement, the procedures of transient detection can be described as the following major parts.

(1) Given a signal $y(t)$ with N data points, find the most similar function $\psi(t, \overline{f}, \overline{\zeta}, \overline{\tau})$ by using correlation filtering based on correlation value maximization rule according to (20)-(21);

(2) construct optimal wavelet basis $\mathrm{Mor}(t, \tau)$ of size $N \times M$ by using the function $\psi(t, \overline{f}, \overline{\zeta}, \overline{\tau})$ above as the basis atom;

(3) construct new transient sparse representation method based on MM algorithm through replacing A with the optimal basis $\mathrm{Mor}(t, \tau)$ in (15); update the new method in (23), and a new data vector $\widehat{c}$ of size $M \times 1$ representing the transients in signal is generated.

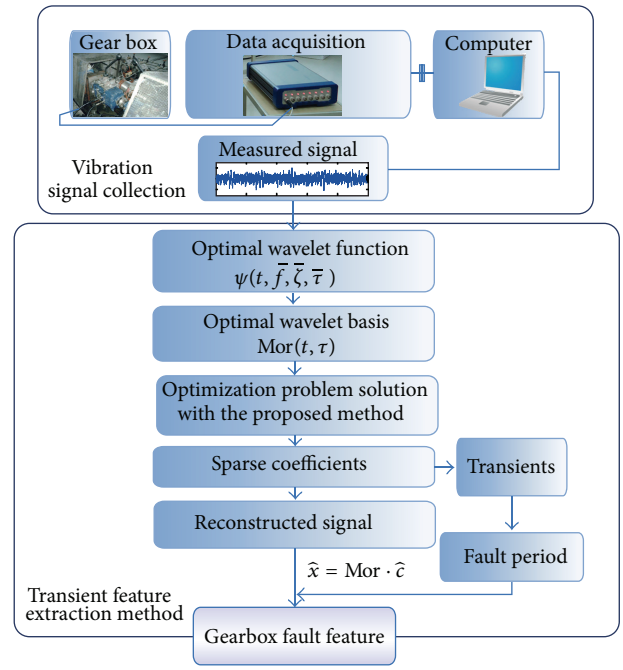


Figure 2: Procedure of the proposed transient sparse representation method.

Finally, the transients hidden in signal can be represented as a sparse vector $\widehat{c}$, and the denoised signal $\widehat{x}$ is finally reconstructed by $\widehat{x} = \mathrm{Mor} \cdot \widehat{c}$.

3. Simulation and Evaluation of the Proposed Method

To verify the effectiveness of the proposed method, a simulated vibration signal $y(t)$ is performed for transient feature extraction. Considering the characteristics of vibration signals of faulty gearbox, the simulated signal is constructed as

$$
\begin{aligned}
y(t) &= x(t) + n(t) \\
&= \sum_k e^{(-\zeta_0 / \sqrt{1-\zeta_0^2})[2\pi f_0 (t - kT_0 - \tau_0)]^2} \\
&\quad \times \cos\left(2\pi f_0 (t - kT_0 - \tau_0)\right) + A_n n(t),
\end{aligned}
\tag{24}
$$

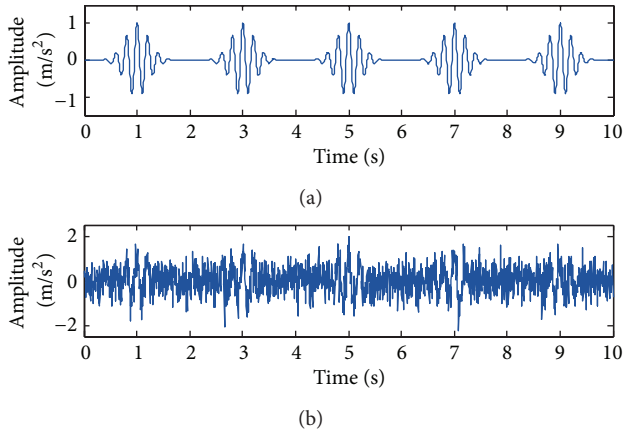


FIGURE 3: The simulated signal: (a) the original signal and (b) the noisy signal.

where frequency is $f_0 = 5$ Hz, $\zeta_0 = 0.01$ is the damping ratio, time index is $\tau_0 = 1$ s, and cyclic period is $T_0 = 2$ s. Obviously, $x(t)$ is a real periodic cyclic impulse response signal simulating the gear fault signal, and the signal $n(t)$ is white noise. The sampling frequency is 200 Hz in time ranges $[0, 10]$ s. Waveform of the simulated signal is shown in Figure 3, whose noise is weighted by $A_n = 0.5$. Figures 3(a) and 3(b) give the waveform of the simulated signal including noise or not, respectively.

The proposed transient sparse representation method is applied to analyze the simulated signal and extract the transients from the signal. According to the procedure in Figure 2, the first step is to calculate the optimal wavelet function by correlation filtering based on correlation value maximization rule. With this principle, parameters $\overline{f} = 5$ Hz, $\overline{\zeta} = 0.01$ are obtained for the optimal wavelet function. The parameters are exactly equal to the simulated values.

The second step is to cope with the objective function as described in (3) (with $\lambda = 8.457$). The optimal wavelet basis $\mathrm{Mor}(t, \tau)$ with parameters $\overline{f}$, $\overline{\zeta}$ in the first step should be founded so as to realize the feature detection sparsely. With the optimal basis $\mathrm{Mor}(t, \tau)$, the transients in the noisy signal are converted into sparse vector $\widehat{c}_{\mathrm{MM}}$ (represents the sparse vector by using MM method) as illustrated in Figure 4(a) through the iterative algorithm shown in (23). The reconstructed signal is shown in Figure 4(c). For better comparison of the sparse vector and the transients in the reconstructed signal, the first N elements of vector $\widehat{c}$ are taken as shown in Figure 4. The following analysis will refer to this one, and not explain in the next section.

In Figure 4(a), cyclic period $T = 2$ s is identified, which is equal to the simulated value ($T_0 = 2$ s). The impulse times are 1 s, 3 s, 5 s, 7 s, and 9 s, respectively, which can also be identified from Figure 4(a). Compared to the original signal shown in Figure 3(a), it can be found that the reconstructed signal shows an excellent effect on transient detection as the transients are captured clearly, while the noise is well discarded.

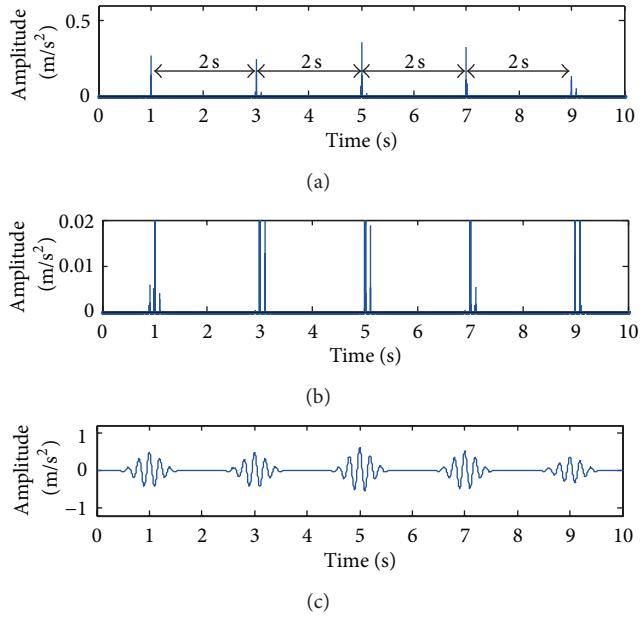


FIGURE 4: Detection results of the simulated signal by using the proposed method: (a) the sparse coefficients; (b) the amplified sparse coefficients, and (c) the reconstructed signal.

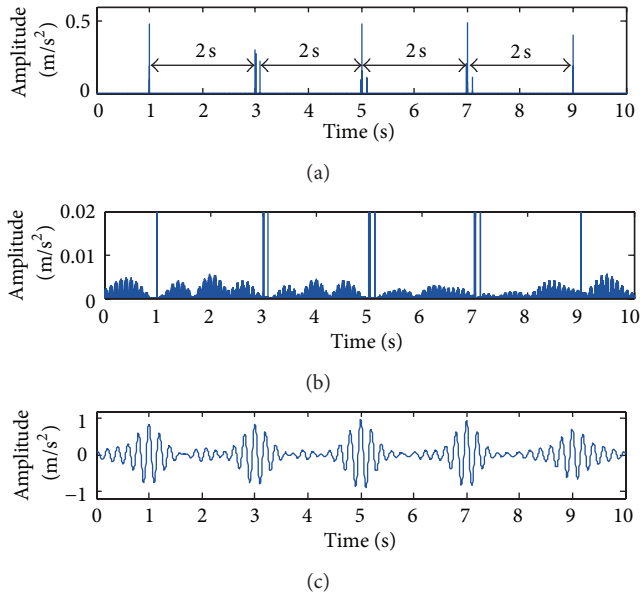


FIGURE 5: Detection results of the simulated signal by using SALSA method: (a) the sparse coefficients; (b) the amplified sparse coefficients, and (c) the reconstructed signal.

To confirm the superiority of this result, the same simulated signal is analyzed by SALSA method for comparison. Optimal Morlet wavelet basis is firstly constructed in the same way described in Section 2.3, and then SALSA algorithm is applied to deal with this detection problem. The detection results of SALSA method are illustrated in Figure 5. Figures 5(a) and 5(c) give the sparse vector $\widehat{c}_{\mathrm{SALSA}}$ (represents the sparse vector by using SALSA) and the reconstructed signal, respectively.

Table 2: The analysis results of the proposed method at different noise amplitudes.

A_n	SNR (dB)	k_r	Impulse time (s)					Period parameter (s)
0	∞	1.000	1.000	3.000	5.000	7.000	9.000	2.000
0.1	10.054	1.000	1.000	3.000	5.000	7.000	9.000	2.000
0.2	4.033	0.998	1.000	3.000	5.000	7.000	9.000	2.000
0.3	0.512	0.997	1.000	3.000	5.000	7.000	9.000	2.000
0.4	−1.987	0.990	1.000	3.000	5.000	7.000	9.000	2.000
0.5	−3.925	0.989	1.000	3.000	5.000	7.000	9.000	2.000
0.6	−5.509	0.984	1.000	3.000	5.000	7.000	9.000	2.000
0.7	−6.848	0.969	1.000	3.000	5.000	7.000	8.995	1.999
0.8	−8.008	0.973	1.000	3.005	5.000	7.000	9.095	2.024
0.9	−9.031	0.965	1.000	3.005	5.000	7.000	9.095	2.024
1.0	−9.946	0.957	1.000	3.005	5.000	7.000	9.095	2.024
1.1	−10.774	0.946	0.995	3.095	5.000	7.000	9.095	2.000
1.2	−11.530	0.933	0.995	3.100	5.000	7.000	9.090	2.024

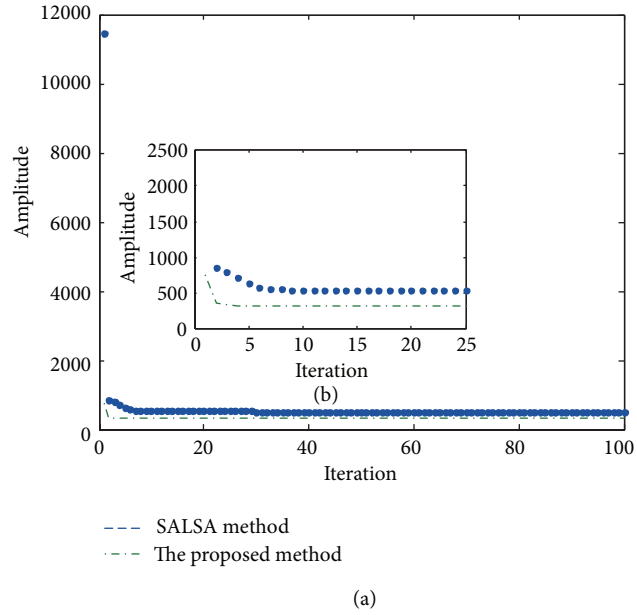
The period of the impulses can also be identified as marked in Figure 5(a) by SALSA method. It seems that SALSA method has a good performance for transient detection. However, some details about the analysis results are ignored. We amplify both the sparse coefficients $\widehat{c}_{\mathrm{MM}}$ and $\widehat{c}_{\mathrm{SALSA}}$ with their y-labels range [0 0.02] as shown in Figures 4(b) and 5(b), respectively. The sparsity of the vector can be measured by the l_0-norm ($\|\cdot\|_0$) of it. We set the threshold value as $1e^{-10}$, then we get $\|\widehat{c}_{\mathrm{MM}}\|_0 = 71$, while we get $\|\widehat{c}_{\mathrm{SALSA}}\|_0 = 1031$. Thus, conclusions can be drawn that vector $\widehat{c}_{\mathrm{MM}}$ is much sparser than vector $\widehat{c}_{\mathrm{SALSA}}$, which implies the proposed method has a better performance on transient sparse representation. Furthermore, their histories of the cost functions are illustrated in Figure 6, simultaneously. History of the cost function shown in (13) of the proposed method is illustrated in green dash-dot line, while the cost function history of the SALSA method [21] for this simulation is illustrated in blue dash line. Figure 6 shows the proposed method has a faster convergence than SALSA method.

Another simulated signal with larger noise ($A_n = 0.7$) as shown in Figure 7(a) is analyzed by these two methods, and the detection results of the proposed method (with $\lambda = 11.840$) and SALSA method are shown in Figures 7(b) and 7(c), respectively. Conclusions can be drawn that, with the increasing of noise level, detection ability of SALSA method fades away and even leads to faulty detection as shown in red circles in Figure 7(c), while the proposed method still has a good performance on extraction effect.

In order to test the noise tolerance of the proposed method, the simulation tests with different noise amplitudes A_n from (24) are investigated as shown in Table 2, in which noise amplitudes A_n, SNR, the correlation coefficients k_γ between the reconstructed signal and the original signal, the impulse time, and the period parameters are listed.

Correlation coefficient k_γ, used to measure the effect of the construction, also can be written as:

$$k_r = \frac{|\langle x, \widehat{x}\rangle|}{\|x\|_2 \|\widehat{x}\|_2}, \tag{25}$$



(a)

Figure 6: The cost function histories of the proposed method and SALSA method.

where x and $\widehat{x}$ are the original component and the corresponding estimate result, respectively.

SNR, the signal-to-noise ratio, used to weigh the noise level is defined as

$$\mathrm{SNR} = 10 \times \lg\left(\frac{P_s}{P_n}\right), \tag{26}$$

where P_s is the energy of the useful information $x(t)$ and P_n is the energy of the noise $n(t)$. As shown in Table 2, it can be seen that with the increasing of the noise amplitude, the proposed method still has an excellent detection effect.

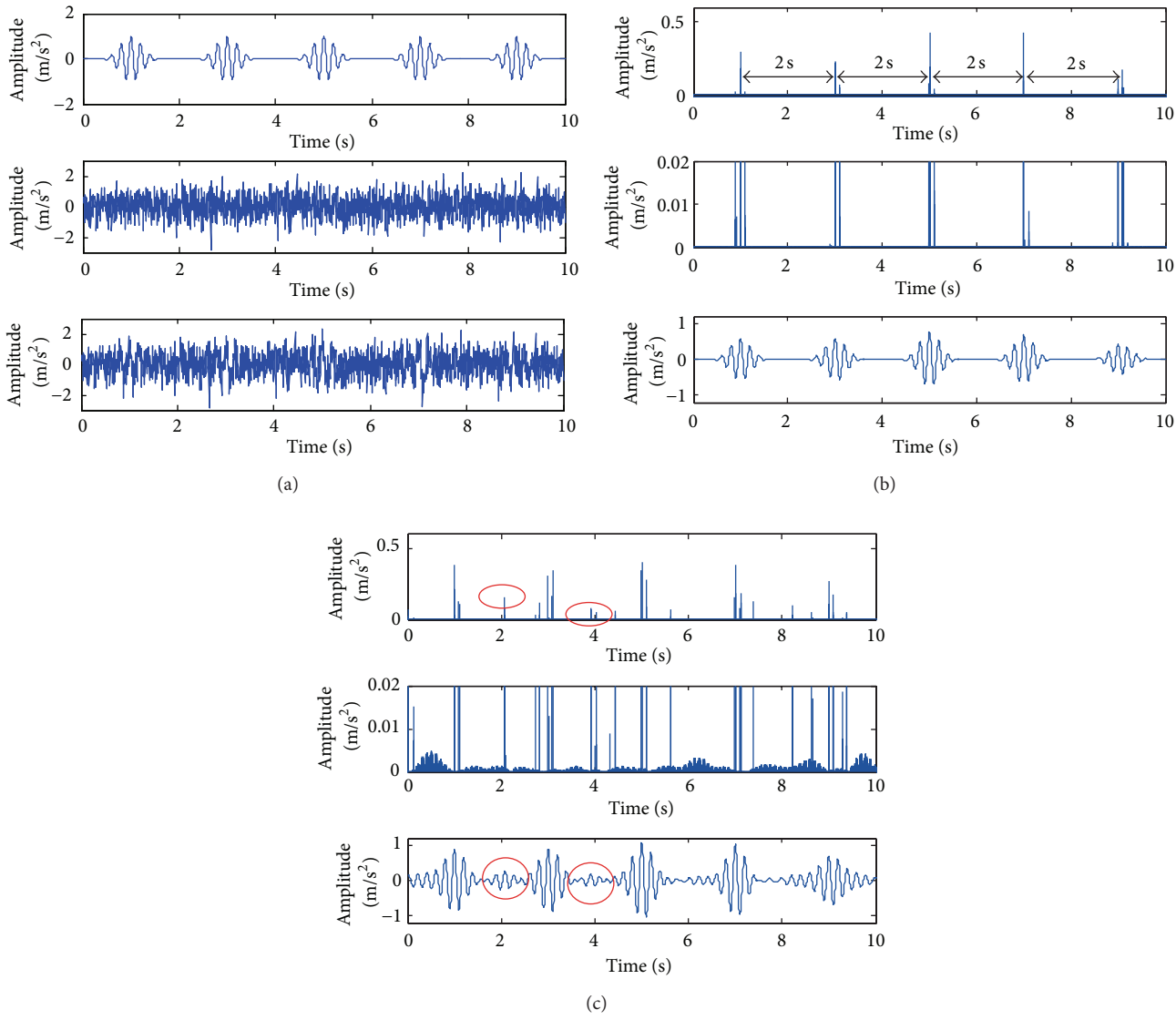


FIGURE 7: Detection results of these two methods: (a) the simulated signal with much noise; (b) the proposed method, and (c) the SALSA method.

4. Experimental Verification

To verify the effectiveness of the proposed method in practical engineering applications, defective gearbox data is analyzed. The experimental data were acquired from an automobile transmission gearbox, which has five forward speeds and one backward speed, as shown in Figure 8. At constant rotating speed, localized faults in gearbox tend to result in periodic shocks and thereby arouse periodic transients in vibration signals, so the transients can be represented as certain wavelet basis sparsely; thus the fault features can be extracted sparsely by the proposed method. To further demonstrate the effectiveness of the proposed method, the results of fault feature detection by the proposed method are compared with the detection results of SALSA method.

During the test, a broken-tooth fault occurred on the driving gear of the third speed. The vibration signal was acquired by an accelerometer mounted on the outer case of the gearbox when it is loaded on the third gearbox. For a gear transmission, the meshing frequency f_m is calculated by

$$f_m = \frac{nz}{60i}, \tag{27}$$

where z is the number of the gear teeth, n is the rotating speed of the input shaft, and i is the transmission ratio. The working parameters are shown in Table 3, the meshing frequency here is 500 Hz, and thus the sampling frequency is set at 3000 Hz.

A measured vibration signal caused by one driving gear teeth broken with a length of 900 and its frequency spectrum are shown in Figures 9(a) and 9(b). From Figure 9(a), the impulse period cannot be identified as the noise corruption; from Figure 9(b), the main frequency component can be identified as 500 Hz.

The proposed transient sparse representation method is then applied to the signal. The optimal wavelet function is first calculated by the correlation filtering rule, and the associated

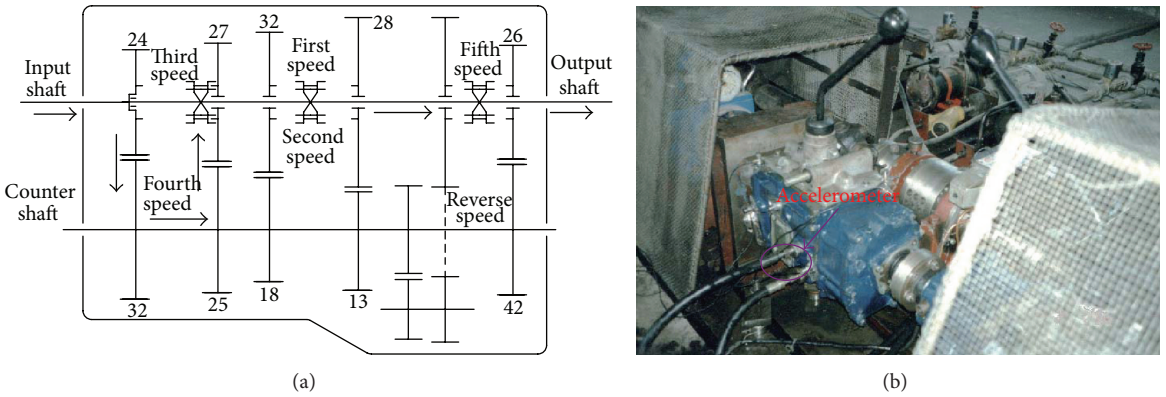


Figure 8: The automobile transmission gearbox: (a) structure of gearbox, and (b) gearbox setup.

Table 3: Working parameters of the third speed gears.

	Number of teeth	Rotating period (s)	Rotating frequency (Hz)	Meshing frequency (Hz)
Driving gear	25	0.050	20	500
Driven gear	27	0.054	18.5	

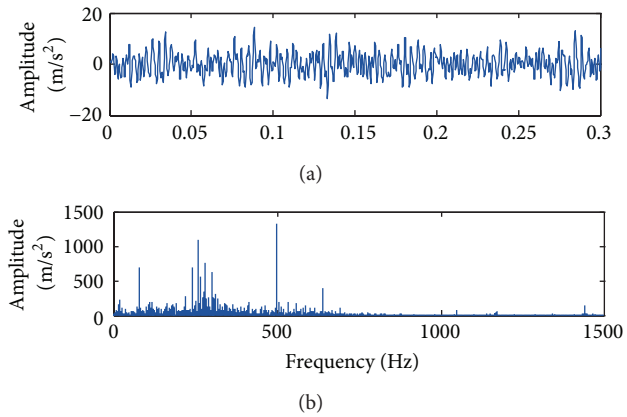


Figure 9: (a) The measured gearbox defective vibration signal, and (b) its Fourier spectrum.

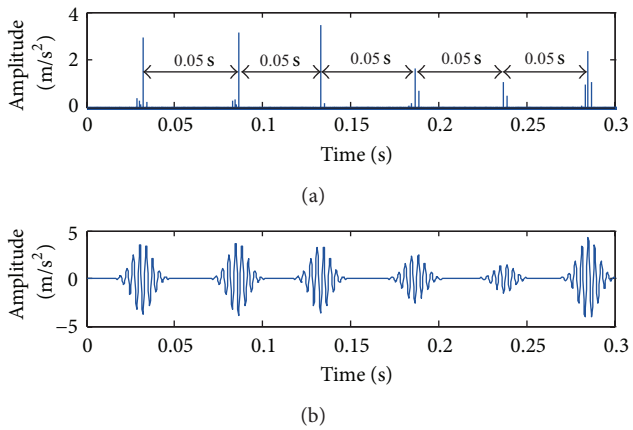


Figure 10: The analysis results of vibration signal by using the proposed method: (a) sparse coefficients, and (b) the reconstructed signal.

parameters are $\overline{f} = 272\,\text{Hz}$,$\overline{\zeta} = 0.0074$, and $\overline{\tau} = 0.0633\,\text{s}$. Then the optimal wavelet basis $\text{Mor}(t, \tau)$ with parameters $\overline{f}$ and $\overline{\zeta}$ is founded so as to realize the feature detection sparsely. Figure 10(a), obtained by the proposed method, gives the sparse coefficients, which represents a series of periodic impulses. The average time period of these impulses is around 0.0505 s, which is very close to the theoretical value of 0.050 s. Figure 10(b) gives the reconstructed signal, whose periodical features represent the localized fault existing in the driving gear of the third speed. The analysis results confirm that the proposed transient sparse representation method can detect the transients clearly and reduce the noise effectively; thus the fault features can be identified.

For the purpose of comparison, the same signal is also processed by using SALSA method, and the results are displayed in Figure 11.

The sparse coefficients obtained by SALSA method are illustrated in Figure 11(a), which theoretically represent a series of periodic impulses. However, from Figure 11(a), the fault period cannot be identified clearly. The reconstructed signal is shown in Figure 11(b). The analysis result in Figure 11(b) has almost no periodical features, thus it is difficult for gearbox fault diagnosis. By comparing with the analysis results in Figure 10, it can be found that the proposed transient sparse representation method is obviously superior to SALSA method.

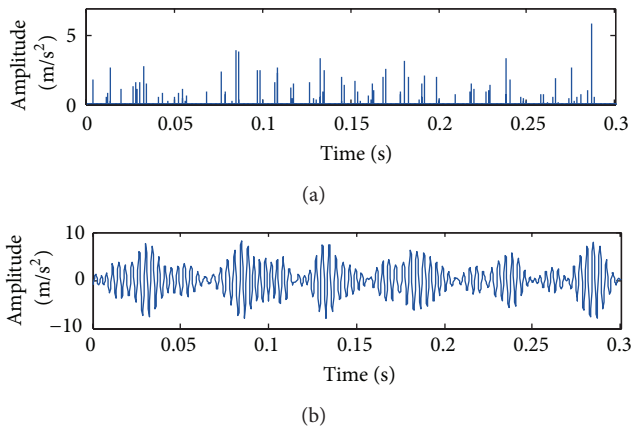


Figure 11: The analysis results of vibration signal by using SALSA method: (a) sparse coefficients, and (b) the reconstructed signal.

5. Conclusions

This paper presents a novel method for transient sparse representation by employing MM algorithm combining wavelet basis to cope with the ill-posed inverse problems. Based on the discussion above, the following conclusions can be drawn.

(1) The proposed method inherits the merits of MM algorithm in noise suppression and sparsity properties to represent the transient features sparsely. Not only is the strong background noise reduced by this method, but also the transients are represented as a series of sparse coefficients, which are of great importance for transient feature detection.

(2) Identification of impulse time and the period parameter based on the proposed method is effective for feature detection, which is demonstrated by the simulated study. Furthermore, the performance of the proposed method has been verified by the application in the defective gearbox data in comparison with the SALSA method. Both the simulation and the application show the proposed method is reasonable and effective in transient detection and thus for machinery fault diagnosis.

Due to the flexibility in the design of the sequence of simpler optimization problems in MM algorithm, the further research may be mainly concentrated on the improvement of sparse algorithm to decrease the computational complexity. Moreover, future research directions will also include repeating the experiments with a wider variety of bases, test signals and other sparsity methods.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of China (nos. 51375322, and 61373098) and also partly supported by the Open Fund of State Key Laboratory for Manufacturing System Engineering (Xi'an Jiaotong University) (no. Sklms2011006).

References

[1] M. Nixon and A. S. Aguado, *Feature Extraction & Image Processing*, Academic Press, Oxford, UK, 2008.

[2] R. X. Gao and R. Q. Yan, "Non-stationary signal processing for bearing health monitoring," *International Journal of Manufacturing Research*, vol. 1, no. 1, pp. 18–40, 2006.

[3] J. Lin and L. S. Qu, "Feature extraction based on morlet wavelet and its application for mechanical fault diagnosis," *Journal of Sound and Vibration*, vol. 234, no. 1, pp. 135–148, 2000.

[4] Z. K. Zhu, R. Q. Yan, L. Luo, Z. H. Feng, and F. R. Kong, "Detection of signal transients based on wavelet and statistics for machine fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 23, no. 4, pp. 1076–1097, 2009.

[5] D. Wang, W. T. Peter, W. Guo, and Q. Miao, "Support vector data description for fusion of multiple health indicators for enhancing gearbox fault diagnosis and prognosis," *Measurement Science and Technology*, vol. 22, no. 2, Article ID 025102, 2011.

[6] H. Liu, J. Wang, and C. Lu, "Rolling bearing fault detection based on the teager energy operator and elman neural network," *Mathematical Problems in Engineering*, vol. 2013, Article ID 498385, 10 pages, 2013.

[7] Y. G. Lei, J. Lin, Z. J. He, and D. Kong, "A method based on multi-sensor data fusion for fault detection of planetary gearboxes," *Sensors*, vol. 12, no. 2, pp. 2005–2017, 2012.

[8] R. Q. Yan, R. X. Gao, and X. F. Chen, "Wavelets for fault diagnosis of rotary machines: a review with applications," *Signal Processing A*, vol. 96, pp. 1–15, 2014.

[9] He, Z. j, Y. Y. Zi, Q. F. Meng, and J. Zhao, *Fault Diagnosis Principle of Non-Stationary Signal and Applications to Mechanical Equipment*, Higher Education Press, Beijing, China, 2001.

[10] Q. B. He, X. X. Wang, and Q. Zhou, "Vibration sensor data denoising using a time-frequency manifold for machinery fault diagnosis," *Sensors*, vol. 14, no. 1, pp. 382–402, 2013.

[11] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics-A tutorial," *Mechanical Systems and Signal Processing*, vol. 25, no. 2, pp. 485–520, 2011.

[12] A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical Systems and Signal Processing*, vol. 20, no. 7, pp. 1483–1510, 2006.

[13] X. F. Chen, X. Li, S. B. Wang, Z. B. Yang, B. Q. Chen, and Z. J. He, "Composite damage detection based on redundant second-generation wavelet transform and fractal dimension tomography algorithm of lamb wave," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 5, pp. 1354–1363, 2013.

[14] S. B. Wang, X. F. Chen, and G. G. Cai, "Matching demodulation transform and synchrosqueezing in time-frequency analysis," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 69–84, 2013.

[15] S. B. Wang, X. F. Chen, G. Y. Li, X. Li, and Z. J. He, "Matching demodulation transform with applications to feature extraction of rotor rub-impact fault," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 5, pp. 1372–1383, 2013.

[16] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.

[17] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[18] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "A fast algorithm for the constrained formulation of compressive image reconstruction and other linear inverse problems," in *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, pp. 4034–4037, March 2010.

[19] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, 2010.

[20] I. Selesnick, "Penalty and shrinkage functions for sparse signal processing," 2014, http://cnx.org/content/m45134.

[21] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2345–2356, 2010.

[22] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2980–2991, 2007.

[23] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[24] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[25] G. Cai, X. Chen, and Z. He, "Sparsity-enabled signal decomposition using tunable Q-factor wavelet transform for fault feature extraction of gearbox," *Mechanical Systems and Signal Processing*, vol. 41, no. 1, pp. 34–53, 2013.

[26] I. Selesnick, "Introduction to sparsity in signal processing," 2014, http://cnx.org/content/m43545/1.3.

[27] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[28] J. M. Bioucas-Dias, M. A. T. Figueiredo, and J. P. Oliveira, "Adaptive total variation image deconvolution: a majorization-minimization approach," in *Proceedings of the 14th European Signal Processing Conference (EUSIPCO '06)*, pp. 1–4, September 2006.

[29] M. A. T. Figueiredo, J. B. Dias, J. P. Oliveira, and R. D. Nowak, "On total variation denoising: a new majorization-minimization algorithm and an experimental comparison with wavalet denoising," in *Proceedings of the 2006 IEEE International Conference on Image Processing (ICIP '06)*, pp. 2633–2636, October 2006.

[30] P. Y. Chen and I. W. Selesnick, "Translation-invariant shrinkage/thresholding of group sparse signals," *Signal Processing*, vol. 94, no. 1, pp. 476–489, 2014.

[31] J. Lin and M. J. Zuo, "Gearbox fault diagnosis using adaptive wavelet filter," *Mechanical Systems and Signal Processing*, vol. 17, no. 6, pp. 1259–1269, 2003.

[32] S. Wang, W. Huang, and Z. K. Zhu, "Transient modeling and parameter identification based on wavelet and correlation filtering for rotating machine fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 25, no. 4, pp. 1299–1320, 2011.

[33] Q. Miao, C. Tang, W. Liang, and M. Pecht, "Health assessment of cooling fan bearings using wavelet-based filtering," *Sensors*, vol. 13, no. 1, pp. 274–291, 2013.

[34] L. C. Freudinger, R. Lind, and M. J. Brenner, "Correlation filtering of modal dynamics using the Laplace wavelet," in *Proceedings of the 1998 16th International Modal Analysis Conference (IMAC '98)*, vol. 2, pp. 868–877, February 1998.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 315901, 17 pages
http://dx.doi.org/10.1155/2014/315901

Research Article

Stochastic Resonance with a Joint Woods-Saxon and Gaussian Potential for Bearing Fault Diagnosis

Haibin Zhang, Qingbo He, Siliang Lu, and Fanrang Kong

Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, Anhui 230026, China

Correspondence should be addressed to Qingbo He; qbhe@ustc.edu.cn

Received 17 April 2014; Revised 11 May 2014; Accepted 16 May 2014; Published 9 June 2014

Academic Editor: Xuefeng Chen

This work aims for a new stochastic resonance (SR) model which performs well in bearing fault diagnosis. Different from the traditional bistable SR system, we realize the SR based on the joint of Woods-Saxon potential (WSP) and Gaussian potential (GP) instead of a reflection-symmetric quartic potential. With this potential model, all the parameters in the Woods-Saxon and Gaussian SR (WSGSR) system are not coupled when compared to the traditional one, so the output signal-to-noise ratio (SNR) can be optimized much more easily by tuning the system parameters. Besides, a smoother potential bottom and steeper potential wall lead to a stable particle motion within each potential well and avoid the unexpected noise. Different from the SR with only WSP which is a monostable system, we improve it into a bistable one as a general form offering a higher SNR and a wider bandwidth. Finally, the proposed model is verified to be outstanding in weak signal detection for bearing fault diagnosis and the strategy offers us a more effective and feasible diagnosis conclusion.

1. Introduction

Stochastic resonance (SR) has been firstly introduced by Benzi and coworkers [1] to explain the more or less periodic occurrence of earth's ice ages. We know the main effect of SR is to enhance the response of a bistable system to weak periodic driving by the injection of random noise [2]. As researchers found that the phenomenon of SR benefits weak signal detection under the background noise, during the past three decades, SR being one of the most exciting nonlinear phenomena has attracted considerable attentions in a wide range of research [3–9]. Via the SR, the output signal of a nonlinear dynamic system can be enhanced with optimized signal-to-noise ratio (SNR) by means of noise addition to the system.

On the other hand, rotating machinery plays a significant role in a wide range of industrial applications, such as transportation vehicles, aeroengine, and power generators [7], while rolling bearing works as a necessary part for rotating machine. Therefore safe and reliable operation of a rolling bearing is an important guarantee to reduce economic losses and avoid personal injury. As a result, accurate health monitoring and diagnosis system is of great significance to indicate the incipient fault that may occur in a rotating machine [10]. In the past decades, many algorithms for this particular application have been proposed [11]. The most common methods are based on signal analysis technique to extract the useful features from the original signals acquired by kinds of sensors. However, the fault signal of the rolling bearing is often overwhelmed with heavy background noise coming from other coupled machine components and working environment, which makes some incipient faults not easy to be recognized. The challenge of fault recognition requires enhancing the weak fault information from heavy background noise.

Noise filtering seems a common and effective method which can suppress the noise and improve the output signal-to-noise ratio (SNR), while the SR offers us a new approach and idea. As compared to traditional techniques that mainly focus on how to suppress the noise, the SR has the merit of signal enhancement by the aid of the noise which means the noise plays an active role in our work [2]. During the development of SR in signal enhancement, many researches focus on some issues referred to next. For the restriction of small parameter, several system parameter tuning or noise intensity tuning methods have been proposed to make them more adaptive, such as frequency shifted and rescaling [3, 4], modulation and demodulation [5], multiscale noise tuning

[7, 8, 12], normalized scale transform [13], and adaptive step-changed [14]. Besides, SR-based signal processing could be improved for a better performance with target of higher SNR or SNRI. Cascaded bistable system was proposed by He et al. [15], which connects two or more bistable systems in series. Li et al. studied the SR effects with multistable potential model [6]. Zhang et al. referred to a multiscale bistable stochastic resonance array [9], while most of current SR researches focus on the traditional bistable potential model with a reflection-symmetric quartic potential. Lu et al. firstly extended the application of Woods-Saxon stochastic resonance (WSSR) which is a monostable SR system to the weak signal detection area [16]. Motivated by this work, we propose a new potential model that maintains the advantages of the WSSR to improve the output SNR and broaden the bandwidth of characteristic frequency and noise intensity.

In this paper, we change the WSSR from monostable into bistable by the cooperation of a Gaussian potential (GP). The WSP was firstly put into use by Deza et al. for wide spectrum energy harvesting based on the SR principle [17]. It is a heuristic potential with three parameters which reproduces the qualitative features of the mean force exerted on each nucleon inside the atomic nucleus. The WSP is monostable, and it will show different shapes with different potential height, width, and wall steepness by tuning its system parameters. Different shapes of WSP in SR system can yield different outputs for the same input signal [16, 17]. Gaussian potentials have been used extensively in nuclear physics as a basis for two-body interactions firstly [18–20]. Later, researchers found its favorable application in solution of differential equation especially the solution of Schrodinger equation. Stephenson applied it to the calculation of the eigenvalues of the three-dimensional Schrodinger equation with an attractive radial GP [21]. Eigenvalues and approximate eigenfunctions of the Schrodinger equation with an attractive radial GP are obtained from a first-order perturbation treatment based on a scaled harmonic oscillator model by Cohen [22]. A GP clearly satisfies the conditions of symmetry, continuity, and confining. This model potential is smooth and possesses a finite depth as well as a characteristic finite radius. As a result of these characteristics, we set a GP working together with WSP where the GP acts as a potential barrier that will change the monostable WSSR into a bistable Woods-Saxon and Gaussian SR (WSGSR) system. The SR effect comes up much more easily with the help of noise and a higher output SNR than we can get.

The rest of the paper is arranged as follows. In Section 2, we introduce the framework of the model of WSP and GP as well as the way they unite together as a bistable model. Section 3 shows us the WSGSR system driven by the joint potential. It gives the criterion for evaluation and the steps to realize weak signal detection based on WSGSR. In Section 4, we utilize the simulated bearing fault signals to evaluate the proposed WSGSR system in comparison with traditional bistable SR, also the WSSR. In Section 5, experimental studies by some practical defective bearing signals are conducted to confirm the effectiveness of the proposed system and method and this reveals the engineering application of the proposed model. Finally, conclusions are drawn in Section 6.

2. Potential Model

2.1. Woods-Saxon Potential. In WSP model, the potential function $U_{\mathrm{WS}}(x)$ is a symmetric nonlinear potential. Its form was proposed in the midfifties by Woods and Saxon [23] and firstly used in the harvest of energy by Deza et al. [17] as a mean-field potential for single-nucleon states within the shell model of nuclear structure which can be expressed as follows:

$$U_{\mathrm{WS}}(x) = -\frac{V_1}{1+\exp\left(\left(|x|-R_1\right)/a\right)}. \tag{1}$$

Here parameter V_1 affects the depth of the potential and parameter R_1 works on the width of the potential while parameter a determines the wall steepness of the potential. This potential with three parameters reproduces the qualitative features of the mean force exerted on each nucleon inside the atomic nucleus. Figure 1 shows us the effects that the three parameters have on the potential distribution with the changing of x. The black curves show the potential $U_{\mathrm{WS}}(x)$ at fixed $V_1 = 2$, $R_1 = 1$, and varying a from 0.02 to 0.2. The result tells us that the value a will affect the plainness of the potential well. It will become flatter with a smaller a and even a square well with width of $2R_1$ for $a = 0$. As a grows the potential walls may become smoother, maintaining the value $-V_1/2$ at $|x| = R_1$. The blue curves indicate the influence of V_1 while the red ones reveal that of R_1. V_1 plays an effect on the depth and R_1 works on the width of the well. Hence the potential shape can be adjusted by parameters V_1, R_1, and a separately or jointly.

2.2. Gaussian Potential. The attractive radial Gaussian potential of the form

$$U_G(x) = -V_2 \exp\left(-\frac{x^2}{R_2^2}\right) \tag{2}$$

is of importance and has been used extensively in nuclear physics during the past decades. As a confining potential, GP has a behavior for large x laid down rapidly [18]. Here V_2 works on the depth of the potential well and R_2 is the range of the confinement potential, which corresponds to the radius. Figure 2 shows us the potential distribution as well as the effects of the parameters. We can find that the potential well with depth of V_2 converges to 0 rapidly on both the side and the smaller R_2, the steeper potential walls.

2.3. Joint Woods-Saxon and Gaussian Potential. With the two introduced models, we proposed a new bistable potential model WSG by the combination of WSP and GP which can be expressed as follows:

$$\begin{aligned} U(x) = U_{\mathrm{WS}}(x) - U_G(x) &= -\frac{V_1}{1+\exp\left(\left(|x|-R_1\right)/a\right)} \\ &\quad + V_2 \exp\left(-\frac{x^2}{R_2^2}\right). \end{aligned} \tag{3}$$

Its potential distribution is shown in Figure 3 with parameters: $V_1 = 2$, $R_1 = 1$, $a = 0.02$, $V_2 = 1$, and $R_2 = 0.05$.

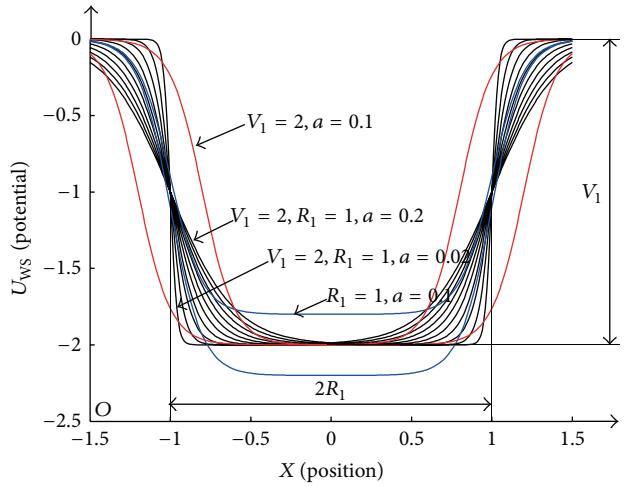


FIGURE 1: Shape of WSP $U_{WS}(x)$ with different parameters. Black curves for different a, blue curves for different V_1, and red curves for different R_1.

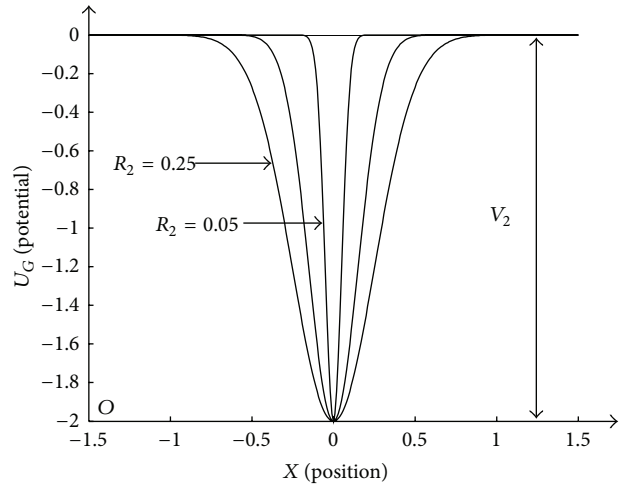


FIGURE 2: Shape of GP $U_G(x)$ with different R_2, where $V_2 = 2$.

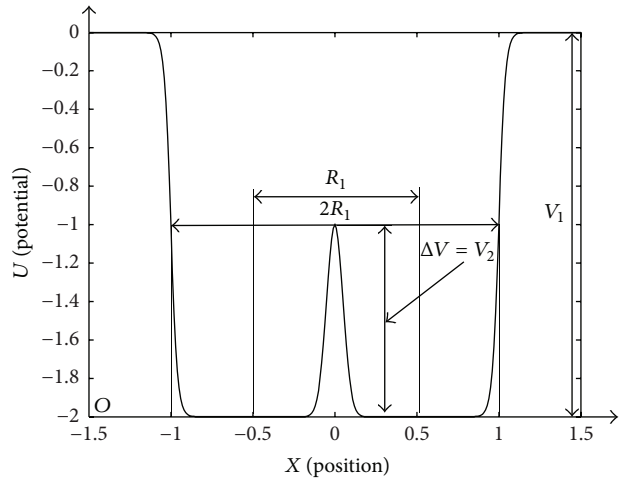


FIGURE 3: Shape of WSGP $U(x)$ with $V_1 = 2$, $R_1 = 1$, $a = 0.02$, $V_2 = 1$, and $R_2 = 0.05$.

We can find a novel bistable potential model where $U(x)$ has two symmetric potential wells with depth of V_2 indicating the barrier height. The width of each well and the distance between the two wells can be adjusted by R_1 and R_2. So the shape of the bistable potential model can be tuned by each parameter independently. The combination of two models works as a mathematical model and the adding of GP makes the monostable WSP turn to a bistable WSGP. The result aids in gaining a new model that contains the superiority of both traditional bistable SR and WSSR. It is still obvious that the potential of the WSG possesses flatter well and steeper barrier whose characteristics will contribute to better performance of SR. Maybe there are some other combinations while they might not offer a bistable model or have the both advantages. Besides, the new model can change easily between the monostable and bistable states, which indicates that (3) works as a more general form. And with this mathematical form, we can process the signal with a novel SR model to extract the weak periodic component that will be introduced later.

3. Stochastic Resonance System

3.1. WSGSR Model. The three basic ingredients of producing SR phenomenon are (1) a bistable or multistable system, (2) a weak coherent input (such as a periodic signal), and (3) a source of noise that is inherent in the system or that adds to the coherent input. Considering the overdamped motion of a Brownian particle in a bistable potential in the presence of noise and periodic forcing, a SR system can be expressed as follows:

$$\frac{dx}{dt} = -V'(x) + A_0 \sin(2\pi f_0 t + \varphi) + n(t), \tag{4}$$

where $V(x)$ denotes the reflection-symmetric quartic potential

$$V(x) = -\frac{a}{2}x^2 + \frac{b}{4}x^4. \tag{5}$$

In (4), we set $n(t) = \sqrt{2D}\xi(t)$ with $\langle n(t)n(t+\tau)\rangle = 2D\delta(t)$, where D is the noise intensity and $\xi(t)$ presents a zero mean unit variance Gaussian white noise. a and b are real parameters. A_0 is the periodic signal amplitude and f_0 is the modulation frequency. In this case, (4) can be written as follows:

$$\frac{dx}{dt} = ax - bx^3 + A_0 \sin(2\pi f_0 t + \varphi) + \sqrt{2D}\xi(t). \tag{6}$$

Equation (6) indicates the nonlinear Langevin equation for one variable x with a traditional reflection-symmetric quartic potential and forms the common bistable SR model. If we change the quartic potential model to the WSGP one, it will be the WSGSR system. Substituting (3) to (4), the new function of SR with WSGP will be in form

$$\frac{dx}{dt} = -\frac{V_1}{a}\operatorname{sgn}(x)\exp\left(\frac{|x|-R_1}{a}\right)\left(1+\exp\left(\frac{|x|-R_1}{a}\right)\right)^{-2} + \frac{2V_2 x}{R_2^2}\exp\left(-\frac{x^2}{R_2^2}\right) + A_0\sin(2\pi f_0 t + \varphi) + n(t), \tag{7}$$

where $\mathrm{sgn}(x)$ denotes the sign function. As the WSGP showed in Figure 3, then what is the difference between the systems indicated by (6) and (7), respectively? It can be seen from (7) that the system output x (left hand side of (7)) is determined by the potential (first two items of right hand side of (7)) and the input signal (third and last items of the right hand side of (7)). The only difference between the two systems is the potential model.

In fact, the effect of SR is the result of the joint action of the system, input signal, and random noise. We build a specific scene that the particle oscillates within the potential under the collective excitation from the potential force, the periodic force, and the noise force to describe SR phenomenon concretely. Among all the kinds of force, the potential force is generated by the gradient of the potential curve which can be described as the first-order derivative of $U(x)$. In practical situation, periodic force and noise force may be constant, so the SR system efficiency will be greatly affected by the potential force to a large extent.

Generally speaking, for the shape of the driving potential such as (3) or (5), if the potential wells are too wide, the particle could not reach the potential wall or the barrier that it always needs to provide the maximal restoring force in one periodic excitation cycle, and hence the restoring force will be helpless in enhancing the particle periodic motion. On the other hand, with too narrow wells, the particle may not have the opportunity to reach the expected destination in the right direction within one excitation cycle because the potential wall may have already forced the particle to move backward prematurely. Similarly, a too high potential barrier will make it difficult for the particle to transmit through the two wells with a periodic input signal of small energy and it may just move among one of the wells. So it cannot work to amplify the small input signal from heavy background noise. But if the barrier height is too low, just the noise force can make the particle traverse without the periodicity. In this case, the particle which means the output oscillates disorderedly and it does not coincide to what we expect. Besides, a steep potential wall will lead to an intense restoring force, which may further cause the particle to be rebounded rapidly. However, when the restoring speed is too high, it may lead to a counteractive effect as the periodic oscillation could not keep pace with the restoring speed. On the contrary, if the potential wall is too gradual, it may not provide enough acceleration to enhance the periodic oscillation. Thus, these indicate that only when the potential is in an optimal condition that can match the periodic force, the periodic and the noise forces could have the effect to amplify the particle oscillation and the periodic signal can then be enhanced.

With these features, we can find the advantages of the WSGSR over the traditional one qualitatively. In order to make the potential model in an optimal status, we need to adjust the parameters. For the traditional potential model as in (5), the parameters a and b work interactively. That is to say, it is not convenient to adjust one isolate potential feature (e.g., barrier height $a^2/4b$) while keeping the other features (e.g., equilibrium positions $\pm\sqrt{a/b}$) invariable by adjusting the parameters. Assuming that the barrier height has been adjusted to the optimal condition, the equilibrium positions or the potential wall steepness may not be in the optimal conditions and still the system is difficult to be tuned to an optimal status. While, for the WSGP as (3) showed in Figure 3, we find the parameters work on the potential features independently, that means we can adjust one of the features to an optimal condition by tuning the parameters without disturbing the others. For example, the barrier height is affected by V_2 only. Note that when $V_2 = 0$, it performs as a WSSR system [16]. So the proposed SR model can be regarded as a common form with WSSR as its specific case. We can also tune a and R_2 to adjust the steepness of the outer and inner side walls, respectively. Parameter V_1 works on the depth and R_1 works on the width of the wells. With these standalone functions, it is quite convenient for us to tune the parameters and make the system gain an optimal status. Besides, it is obvious that the WSGP can form a much steeper walls and flatter well bottoms. This seems to make the particle oscillate more easily. In the traditional potential model, the value of potential may be boundless with a position x large enough. In this circumstance, the system output may be divergent with a large input signal or noise which exceeds the threshold. So the SR effect will not occur. But for the proposed model, it has an upper limit of potential value of V_1. So it is more probable for the output to be convergent and has a wider input amplitude bandwidth. This indicates the SR effect comes up much more easily which will generate a desired output. In our next work, all the description of advantages for new model will be verified by both simulated and experimental results.

3.2. Output Evaluation and Diagnosis Scheme. To optimally detect the weak signal of driving frequency from the background noise, the WSGSR can be conducted by adjusting the WSGP parameters until the optimal SR output signal which we get with the method mentioned above is achieved. While this is not enough, we need a target to evaluate the performance of SR effect. Here, we employ the output SNR as a criterion to assess the result with which we can judge the optimal condition of our SR system. The output SNR is defined as the power spectral density of the driving signal divided by the average background noise in a small frequency bin around the driving frequency [9]. In order to have a simple computational process, we simplify the definition and expression of SNR into the form [6, 8]

$$\mathrm{SNR} = 10\log_{10}\frac{A_d}{A_n}, \tag{8}$$

where A_d and A_n are the amplitude values corresponding to the driving frequency f_d and the strongest interference frequency f_n (which means the frequency with the largest amplitude except the driving frequency) in the power spectrum, respectively. A higher SNR implies a better discrimination between the periodic signal and the noise. The proposed model and the signal processing also aim at gaining a higher SNR.

The SR model is shown as in (4). In the signal processing, it mainly refers to the numerical analysis. In practical application, the periodic signal and noise are all definite with a known potential model. So we need to solve this differential

equation to gain its output $x(t)$. Equation (4) is a typical first-order differential equation, so the discrete fourth-order Runge-Kutta method [16] is used here. With this method, we can get the output easily.

Then based on the new WSGSR model, we propose a new scheme of bearing fault diagnosis with this SR system, which is presented in Figure 4. In practical applications, the vibration or acoustic signal is usually acquired by accelerometers or microphones attached from bearings which is always modulated to a high frequency band. So in the proposed scheme, the envelope of the analyzed signal is firstly achieved based on the Hilbert transform (HT). Secondly, the modified signal is sent to the WSGSR system in (7) with presetting parameters. In next step, we calculate the output SNR of each set of parameters in the parameter searching space which is used to obtain the optimal parameters corresponding to the maximal SNR. Substitute the optimal parameters to the model and calculate the finally output, then go ahead with the spectral analysis in the spectrum of the system output by Fourier transform, and realize the fault frequency identification.

4. System Performance with Simulated Signal

4.1. Intuitional Output of WSGSR. To have a general presentation for the effect of the proposed SR model, we simulate a sinusoidal signal with frequency of f_d and amplitude of A surrounded with Gaussian white noise of intensity $\sqrt{2D}$

$$s(k) = A\sin\left(2\pi\frac{f_d}{f_s}k\right) + \sqrt{2D}\xi\left(\frac{k}{f_s}\right). \tag{9}$$

Firstly, we take $A = 0.5$ and $f_d = 100$ Hz. The sampling frequency f_s is taken as 20 kHz here. Figures 5(a) and 5(b) are the sinusoidal signal and the mixed signal with noise intensity of 1.5 in time domain, respectively. In Figure 5(b) we can hardly find the periodic component. Figure 5(c) is the potential model of traditional bistable SR as in (5) while Figure 5(d) is the potential model of WSGSR as in (3). All the initial parameters are set as shown in the figures to guarantee a similar potential feature between the two models. Both the barrier heights are 0.25 and the balance positions are set on $x = \pm 1$ which indicate the distance between the two potentials. The similar features can offer us the fairness in comparison of the ability and advantages between the different models. Since we have both the input signal and the SR model as in (6) or (7), we will try the numerical analysis next.

As the power spectral density of SR output meets the Lorentzian distribution which is characterized by concentrating most of noise energy into the low frequency region. So there is a small parameter limitation for SR effect [2]. In practical signal processing, the sampled signal parameters such as the frequency and amplitude of a periodic signal usually exceed the limitation of small parameter SR. In order to normalize the frequency of the original signal to be a small one, the rescaling frequency method is employed here [3]. We take the rescaling ratio $R = 1000$ which changes the characteristic frequency from 100 Hz to 0.1 Hz that is smaller than one. With a noise intensity of 1.5 and 8192 points data length (all the numerical analysis in this paper is under the same condition), Figures 5(e) and 5(f) indicate the outputs of the two different models, respectively. We can find that, under the same condition, the WSGSR system makes the particle transit between the two potential wells more easily than the traditional system. We have given the explanation before as the potential walls are steeper and this causes the particle to be rebounded rapidly gaining a higher speed. With this advantage, the transition between the two wells comes up much more easily in the proposed SR system which will make the SR effect more likely to be available. Then, keeping all the parameters except the noise intensity, we increase it to 3 and recover the output as Figures 5(g) and 5(h). As a result, the particle oscillates between the potential wells more severely and the transition comes up more obviously. This fits the SR effect well and the transition in Figure 5(h) is more coherent. It profits from the fact that the potential wells in WSGSR are flatter so when the particle locates in one of them, it moves unobstructed. Further on, we reduce the characteristic frequency to 50 Hz while keeping the same rescaling ratio and the rescaling frequency will be 0.05 Hz. In this condition, the two systems' outputs are shown as in Figures 5(i) and 5(j). As expected, we get the output more approximate to the original signal with its waveform highly restored. We still can find a better result of WSGSR which has 15 peaks in 0.3 seconds distinctly. Next step, we try to find the optimal parameters for a certain simulated bearing fault signal that is periodic unilateral attenuation impulse and reveal the proposed model's deeper performance.

4.2. Optimal Output with Simulated Bearing Fault Signal. In practical bearing fault diagnosis, bearing fault signal is always present in the form of impactive series with a fault frequency modulated to a high level [24]. So a periodic unilateral attenuation impulse discrete signal with Gaussian white noise is selected as the simulated bearing fault signal according to the equation

$$\begin{aligned} s(k) = {} & A\sin\left(2\pi\frac{f_m}{f_s}k\right)\cdot\exp\left(-d\cdot\operatorname{mod}\left(f_0k, f_s\right)\right) \\ & + \sqrt{2D}\xi\left(\frac{k}{f_s}\right). \end{aligned} \tag{10}$$

Here $\operatorname{mod}(a, b)$ is the remainder of a divided by b which controls the impulses that appear periodically. $A = 1$ means the amplitude and $f_m = 1200$ Hz indicates the modulation frequency of modulating signal. $f_s = 20$ kHz means the sampling frequency. $d = 600$ reflects the attenuation rate and we take the $f_0 = 100$ Hz here as the fault frequency or driving frequency. $\xi(t)$ is the noise with zero mean and a unit variance with noise intensity $\sqrt{2D} = 1.8$. We take a simulation with 8192 points which means the k changes from 1 to 8192. The generated signals without and with the noise are shown as two figures in Figure 6(a), respectively. Figure 6(b) shows its envelope signal and power spectrum. As a modulated signal, we need to have it demodulated and we use HT method referred to in 3.2 (note that the envelope signals in the following cases are all processed in the same way). Before the demodulation process, filtering of noises

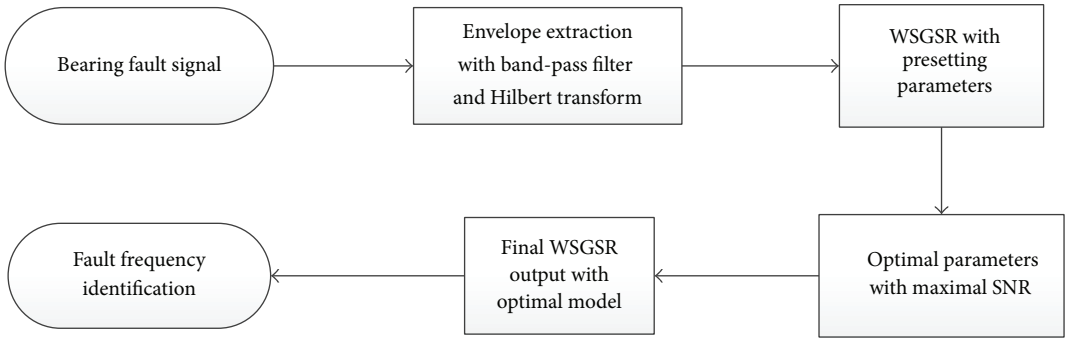


Figure 4: Proposed scheme of bearing fault diagnosis with WSGSR.

that are located in the uncorrelated bandwidth is taken. We can find the signal's main energy band from the spectrum directly and here a four-order Butterworth band-pass filter with bandwidth of 1200 Hz centered at 1200 Hz is used. We can find that the fault frequency f_0 is extracted successfully with HT and band-pass filtering while interferences around f_0 can be seen clearly which makes it hard to identify.

Sending the above signal demodulated by HT into the traditional bistable SR model, we get the output signal and its power spectrum as in Figure 6(c). The parameters are changed to $a = 1$ and $b = 16$ to guarantee an easy appearance of SR which might not be the optimal ones with rescaling ratio of 1000. Pay attention to that the potential barrier height is now $a^2/4b = 1/64$ with equilibrium positions of $\pm\sqrt{a/b} = \pm 0.25$. From the output we can find the component of f_0 is highlighted in the spectrum while the noise components are still frustrating to some extent. What is more is that the output signal in time domain is not satisfying which means the particle cannot transmit between the wells smoothly. Maybe it can be solved by the optimization of parameters. Later we take the WSGSR model with the similar features and the parameters are fixed to match the features of traditional one, where $V_1 = 3$, $R_1 = 0.5$, $a = 0.4$, $V_2 = 1/64$, and $R_2 = 0.4$. We can calculate the two models have the same barrier height and equilibrium positions. The output of the proposed model with the same input signal is shown in Figure 6(d) as well as its spectrum. It can be seen from the power spectrum that the amplitude of component f_0 is higher than that of the other noise components which makes it easily recognized. Besides, the components of lower frequency are weaker than that of the output in traditional SR. So we can find the signal of WSGSR in time domain more ideal which is to say the particles in traditional one move harder as a result of too many low frequency components. It indicates that the particles in new SR system transmit more easily between the two potential wells with the same barrier height and equilibrium position. As the differences between the two models are shape of walls and bottoms, this may owe to the steeper walls and flatter well bottoms which we have described before. The results also show us the superiority of WSGSR over the traditional SR system.

The outputs we get in Figure 6 are not the optimal results as the parameters are preset. So we need to find out a set of parameters to gain the highest SNR as in (8) and an optimal output. Here SNR acts as the criterion. We can take an iterative algorithm to find each optimal parameter within a certain range. After this work we can get the optimal system to match the input signal. The specific steps are introduced as below.

(1) Before we search for the optimal output, we need to have the signal preprocessed. As the original signal is modulated, we firstly need to have it band-pass filtered. Then use Hilbert transform to process the signal. The output will be demodulated successfully. After this, we can send the signal into WSGSR system of different parameters to gain the output.

(2) Calculate the power spectrum of the output waveform and obtain the SNR. Search the maximal SNR in the parameter space that is constructed by the varying variables V_1, R_1, a, V_2, and R_2 in certain ranges and then obtain the optimal parameters corresponding to the maximal SNR. If all the optimal parameters are not on the boundaries of the search ranges, we have found the optimal result successfully or else extend the range and try the searching step again.

(3) Finally, substitute the optimal parameters (as we get in Figure 7 $V_1 = 9$, $R_1 = 0.6$, $a = 0.5$, $V_2 = 0.15$, and $R_2 = 0.9$) and get the output waveform as the detected weak signal. Further calculate the corresponding power spectrum for identification of the driving frequency and make a diagnostic conclusion. Figure 7 shows us the finally optimal output waveform and its spectrum with a clearly highlighted fault frequency or driving frequency of 100 Hz. When comparing it to the initial output as in Figure 6(d), we can find less noise around the driving frequency and a more distinct result which verifies the effectiveness of the optimization process.

4.3. Performance Analysis and Comparison. In the optimization process above, we mentioned that optimal parameters can be obtained according to the maximal SNR in a certain range. It means that there will appear a peak for SNR with only one of the parameters changing. With a decided input signal, the system output only depends on five potential parameters. We fixed four of them and varied the other one of the WSG potential in turn and then calculated the SNR of the output signal via (7) with input signal same as last section

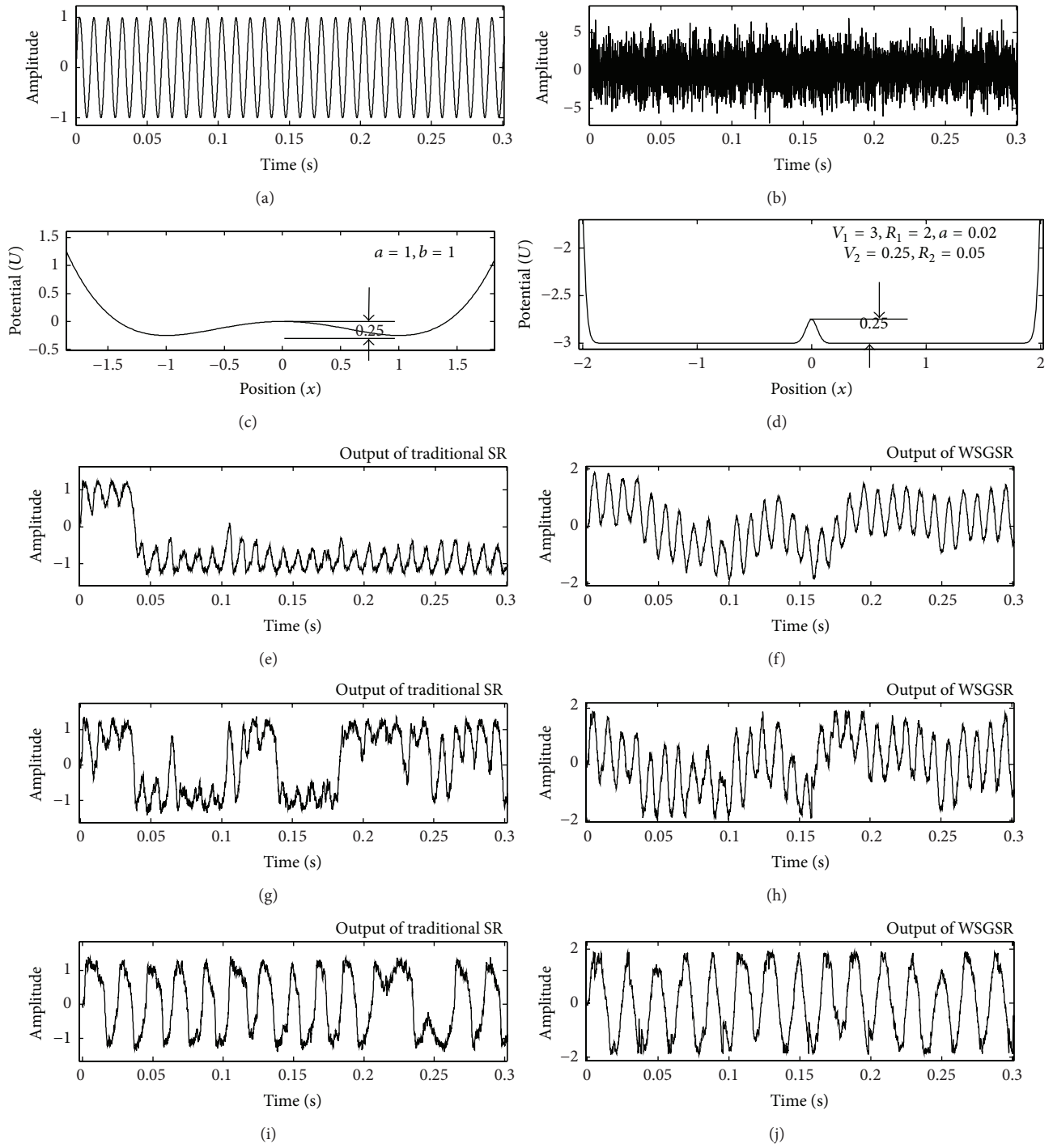


FIGURE 5: Outputs signal of different SR models with a sinusoidal signal mixed with noise: (a) sinusoidal signal, (b) original signal mixed with noise, (c) traditional bistable potential with $a = 1$, $b = 1$, (d) WSGP with $V_1 = 3$, $R_1 = 2$, $a = 0.02$, $V_2 = 0.25$, and $R_2 = 0.05$, (e) and (f) outputs of the two different models with noise intensity of 1.5 and driving frequency of 100 Hz, (g) and (h) outputs of the two different models with noise intensity of 3 and driving frequency of 100 Hz, and (i) and (j) outputs of the two different models with noise intensity of 1.5 and driving frequency of 50 Hz.

whose driving frequency is 100 Hz and noise intensity of 1.8. Figure 8 shows us the SNR variation trend with the changing of five parameters. The thick red curve is the smoothing fitted curve (via piecewise polynomial interpolation method) of the original SNR curve (thin one). In every figure, we vary sequentially each one of the parameters while keeping the others fixed. For example, in Figure 8(a), we make $R_1 = 0.5$, $a = 0.2$, $V_2 = 0.3$, and $R_2 = 0.8$ while varying $V_1 \in [0.05, 10]$. The results offer us the hints that the proposed model has a similar feature as traditional SR system in which the SNR

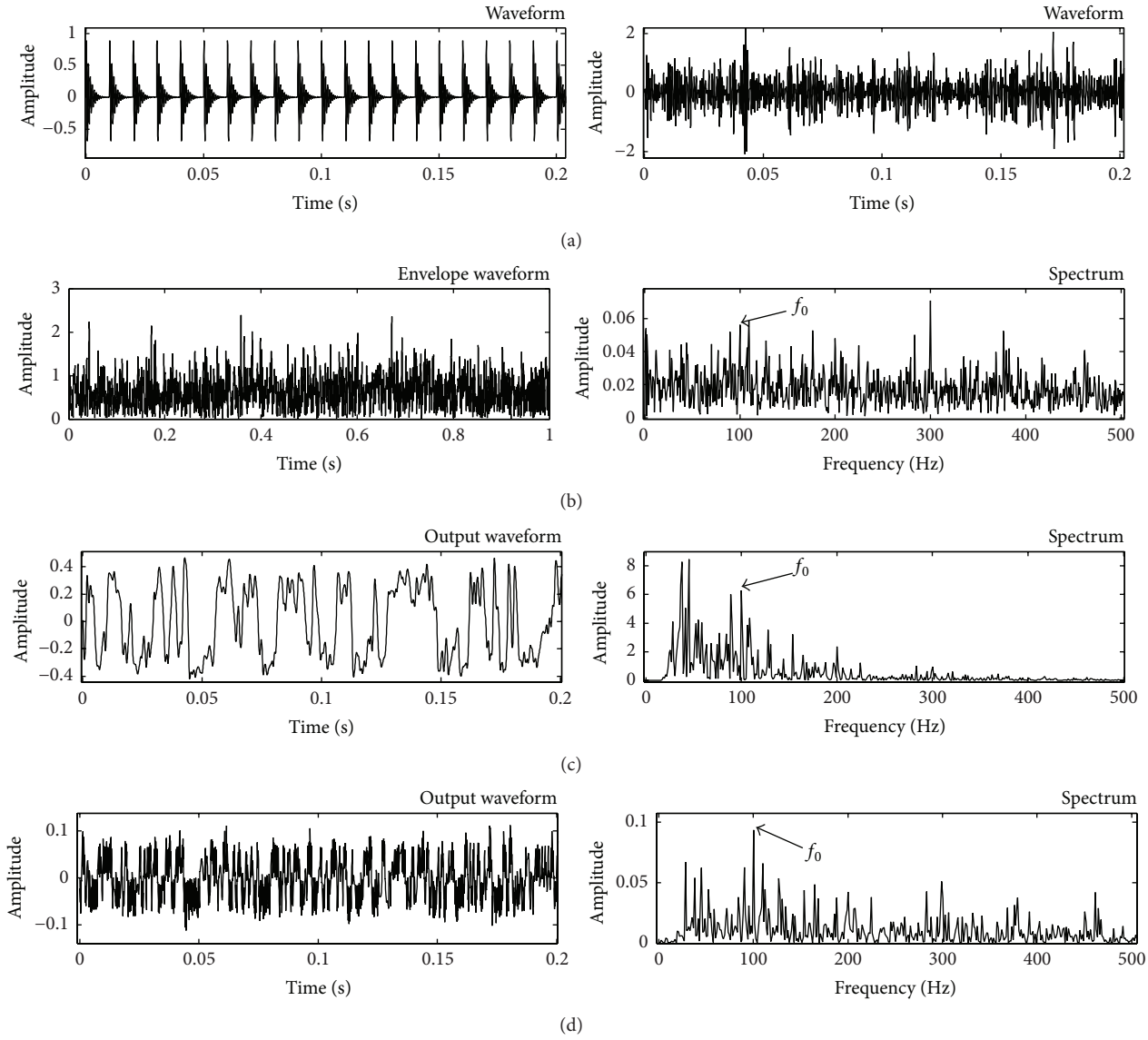


FIGURE 6: Simulated bearing fault signal and outputs of different models: (a) periodic unilateral attenuation impulse signal of 100 Hz and mixed signal with noise intensity of 1.8, (b) envelope signal and power spectrum filtered with [600 Hz, 1800 Hz], (c) output of traditional bistable SR model and its power spectrum with $a = 1$, $b = 16$, and (d) output of WSGSR model and its power spectrum with $V_1 = 3$, $R_1 = 0.5$, $a = 0.4$, $V_2 = 1/64$, and $R_2 = 0.4$.

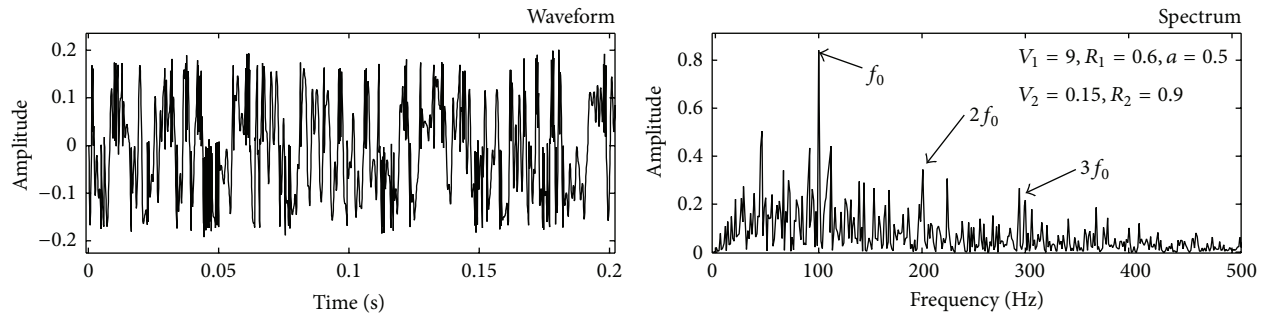


FIGURE 7: Optimal result for Figure 6(d) and its spectrum with $V_1 = 9$, $R_1 = 0.6$, $a = 0.5$, $V_2 = 0.15$, and $R_2 = 0.9$.

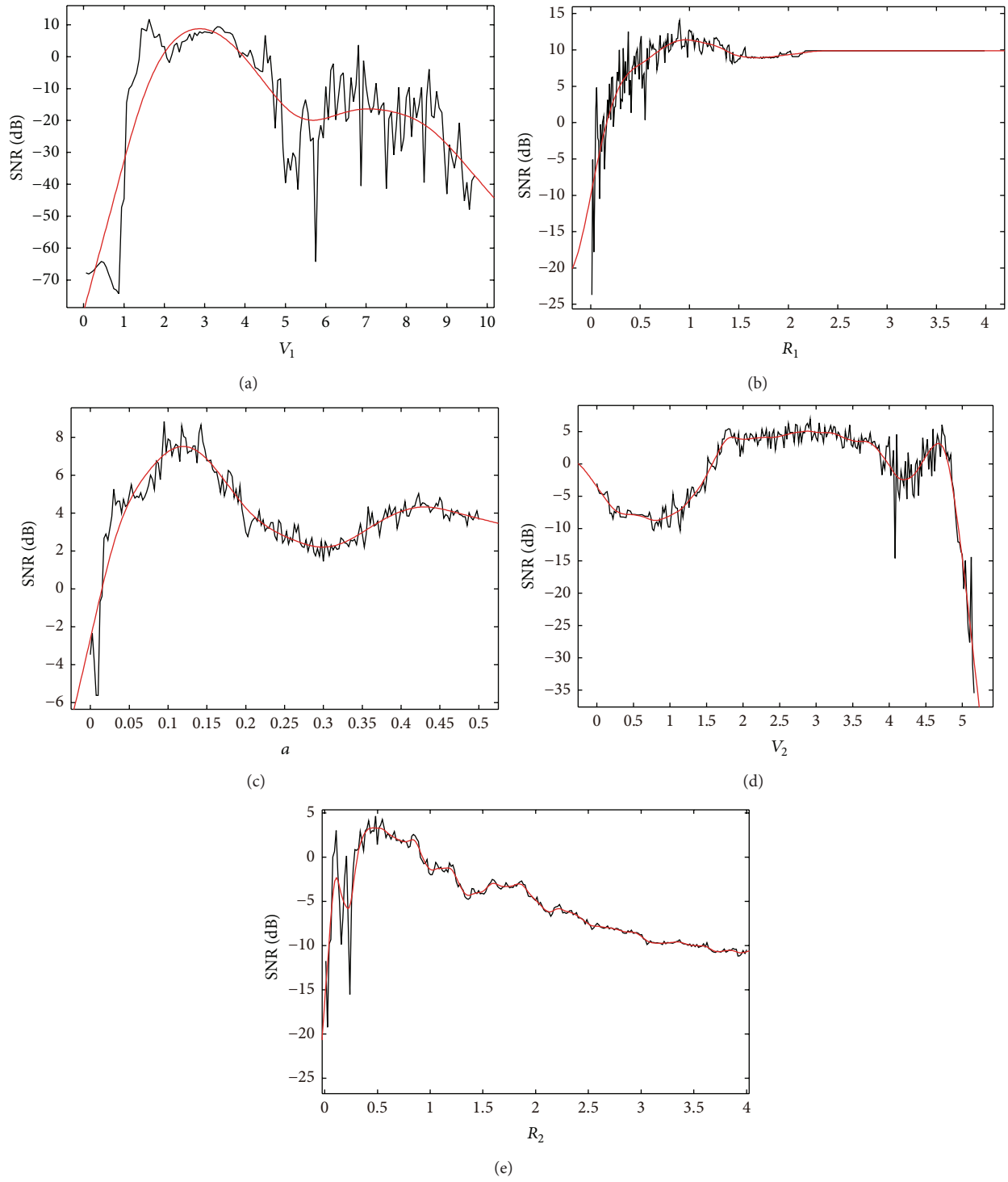


FIGURE 8: SNRs variation trend with the changing of five parameters in different ranges: (a) [0.05, 10], (b) [0.01, 4], (c) [0.1, 0.5], (d) [0, 5.5], and (e) [0.01, 4].

variation tendency firstly increases with the increasing of the variable reaches a maximum and then decreases when the variable keeps increasing. Figures 8(d) and 8(e) especially offer us the evidence that the proposed method can improve the output SNR of WSSR, where $V_2 = 0$ further. Because, in our method, even if we fix $V_2 = 0$ firstly, we can get an optimal output of WSSR model. Then we can tune the parameters V_2 and R_2 to get a higher SNR and a better output signal via these two figures. That is to say, an optimal output with the highest SNR of WSGSR can be obtained by parameter tuning. Besides, with the increasing of V_2 from 0, we find that the SNR decreases firstly. Later, with the changing of V_2 further, the

output SNR begins to increase and even exceeds the situation under $V_2 = 0$. This can be explained easily. Then the adding of V_2 in the beginning plays a negative role to the output. But with the changing of its value, when the system comes to a best matching, the barrier will serve a positive function. It means that the employment of V_2 does not always have benefit unless it is under an optimal situation.

Following on, we make a new simulated signal with a higher driving frequency of $f_0 = 200$ Hz whose amplitude is fixed at $A = 0.8$ and a noise intensity of $\sqrt{2D} = 1.5$ in (10). In Figure 9(a), it stands for the original waveform and its power spectrum. All the other parameters are the same as in Section 4.2. From its spectrum in Figure 9(a), we can easily get the signal's main energy band and design the band-pass filter. Figure 9(b) shows the envelope signal generated by HT and its spectrum after a filter with bandwidth of 1600 Hz centered at 1200 Hz. Without any process except for the filtering and HT, the driving frequency can be hardly identified. Sending the envelope signal into the traditional bistable SR model, optimizing the parameters a and b, we can get a best combination of parameters of 0.2 and 46, respectively, with the rescaling ratio $R = 1000$. Its output is shown as in Figure 9(c) with SNR of 2.89 dB. We find that the driving frequency at 200 Hz is well extracted while there are still too many noises of low frequency that obstruct our judgment. Later, the proposed WSGSR method is utilized to deal with the filtered envelope signal and after parameters tuning we get the optimal values for parameters: $V_1 = 14$, $R_1 = 0.2$, $a = 0.4$, $V_2 = 0.7$, and $R_2 = 0.5$. The optimal output in Figure 9(e) shows us a distinct characteristic frequency with noise highly weakened and the maximal SNR reaches 4.71 dB. Specially, we set $V_2 = 0$ when the model becomes WSSR; we get the optimal parameters combination of $V_1 = 8$, $R_1 = 0.45$, and $a = 0.7$ and its output is shown in Figure 9(d). It is also a satisfying result with SNR of 4.02 dB. Compared with the result of traditional model, the WSGSR output can keep the original waveform better and offer a larger enhancement of the weak impulse signal. Besides, the differences between Figures 9(d) and 9(e) show us the probability that the proposed model may have a better performance which contains the situation of WSSR and can reach a higher output SNR. The amplitude of the driving frequency in its output gets 4.008 while it is only 0.524 when $V_2 = 0$. This means higher output energy with the help of potential barrier like in a traditional bistable SR with amplitude of 1.791 although the SNR is the lowest.

In order to verify our conjecture about the proposed method, we make a simulation to compare the capacity of different models when dealing with signal under different noise intensity or driving frequency.

Firstly, fixing the frequency at 200 Hz still, we set the amplitude of attenuation impulse signal in (10) $A = 0.8$ and the noise intensity $\sqrt{2D}$ varies from 1.1 to 2.3 with step of 0.2 with data length of 10000 points. Then under interference of each noise intensity scale, we can gain every optimal output with a maximal SNR. The output SNR is shown as in Figure 10(a) with the changing of noise intensity. The three different results are yielded by the traditional bistable SR method, the WSGSR method with $V_2 = 0$ (WSSR), and the proposed method, respectively. What is obvious is that the three curves all indicate a decrease of SNR with the increase of noise intensity. This confirms our cognition well. But there is several other information which the figure offers us. We can find a higher SNR with the proposed method than the traditional SR and the employment of V_2 which improves the performance of WSSR offers us a better result. When treating the WSSR as a special case of WSGSR, the proposed model shows the superiority in detecting weak impulse signals that are submerged in noise of different level. Besides, when comparing the distinction between them, we find an increasing trend (from 0.19 dB to 0.81 dB) with the increasing noise intensity which means the heavier noise, the more superior performance the proposed model has relatively.

Secondly, to observe the performances of different models under different driving frequency, we make another series signal. In them, the amplitude of attenuation impulse signal is set as $A = 0.8$ with noise intensity of 1.5. They have varying driving frequency from 50 Hz to 1000 Hz with 6 points. To research the effectiveness and bearing capacity to different driving frequency, we still calculate the optimal output of signal under each driving frequency and get the output SNR. Making a permanent rescaling ratio of 1000, the three curves in Figure 10(b) show us the results. All the three SNRs increase firstly because the higher driving frequency, the higher energy it will contain under same amplitude as (10) gives. We can find a more stable performance with our proposed model than the traditional bistable SR with the changing of driving frequency and the parameter V_2 still works. Although the larger frequency impacts the output SNR, it is not as serious as the traditional bistable SR. It coincides with the result in Section 4.1 well. It is obvious that as a restriction of small parameters, the equivalent driving frequency of traditional bistable SR changes from 0.05 to 1 with rescaling ratio of 1000 which gets farther and farther from the condition. So as a result, with a frequency large enough, the output SNR begins to decrease rapidly. In our proposed model, the SNR also decreases after a certain frequency but not as severe as it. This means the proposed model has an advantage when dealing with high frequency signal over the traditional one.

Based on the above analysis, our conjecture is verified. With the proposed WSGSR model, signal extract under heavy noise or bearing fault diagnosis can be obtained by tuning the independent parameters and it is proved to have a better adaptability in detecting the weak signal with different noise intensities and different driving frequencies. After these simulation works, the model will be applied in experimental bearing fault signal to value its engineering application.

5. Engineering Application

To verify the effectiveness and efficiency of the proposed method in engineering applications, a set of train bearing signals carrying fault information are analyzed according to the scheme of bearing fault diagnosis with proposed WSGSR model in Figure 4. The bearing outer-race defective and inner-race defective datum are generated from a rolling bearing used separately.

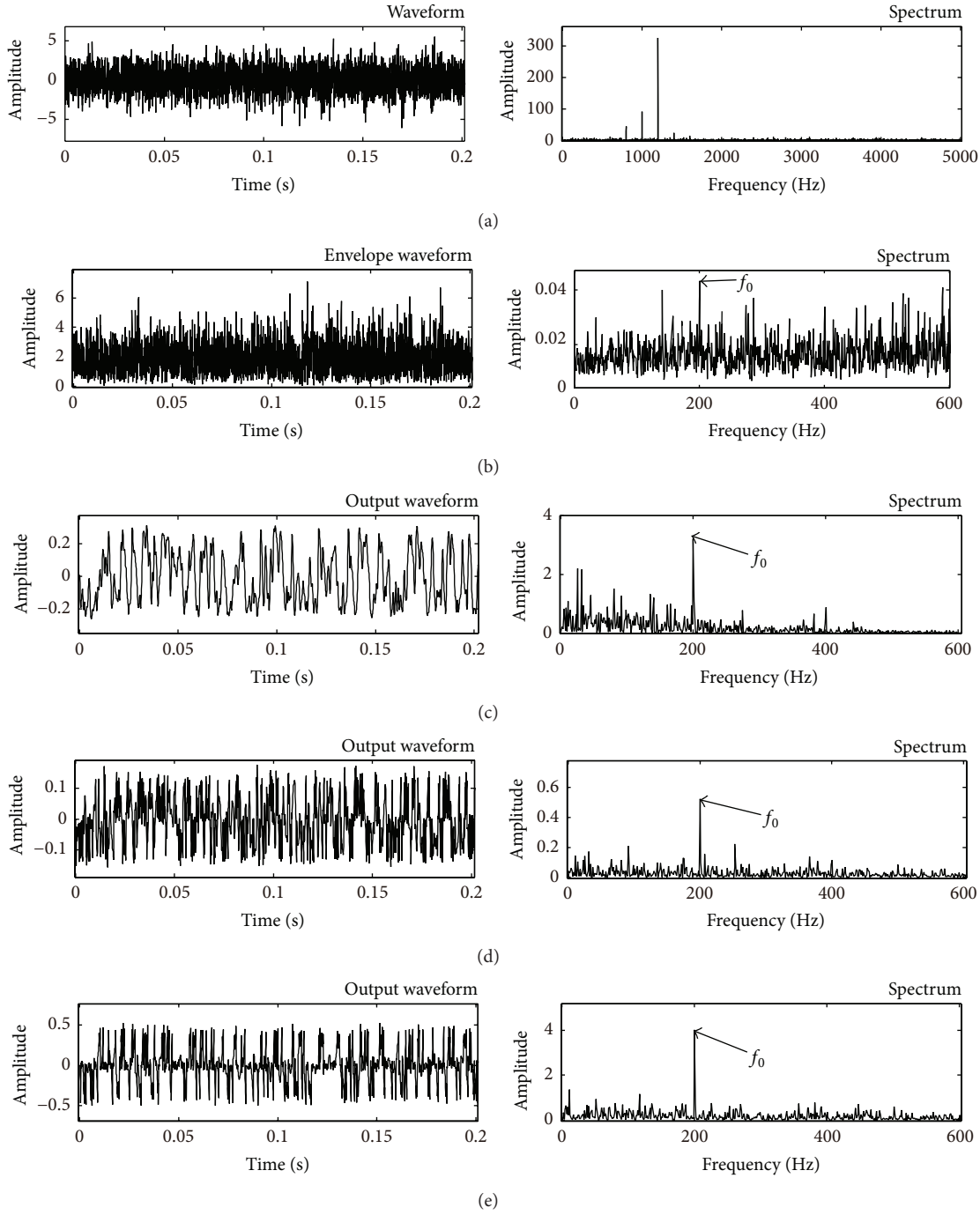


Figure 9: Outputs of different models with driving frequency of 200 Hz: (a) original waveform and its power spectrum with $A = 0.8$, $\sqrt{2D} = 1.5$, (b) envelope signal and its spectrum filtered with [800 Hz and 1200 Hz], (c) output of bistable SR model and its spectrum with $a = 0.2$, $b = 46$, (d) output of WSGSR model and its power spectrum with $V_1 = 8$, $R_1 = 0.45$, $a = 0.7$, and $V_2 = 0$ (WSSR), and (e) output of WSGSR model and its power spectrum with $V_1 = 14$, $R_1 = 0.2$, $a = 0.4$, $V_2 = 0.7$, and $R_2 = 0.5$.

Table 1: Specification of the train bearing NJ(P)3226X1.

Type	Diameter of the outer race	Diameter of the inner race	Pitch diameter (D)	Diameter of the roller (d)	Number of the roller (z)
NJ(P)3226X1	250 mm	130 mm	190 mm	32 mm	14

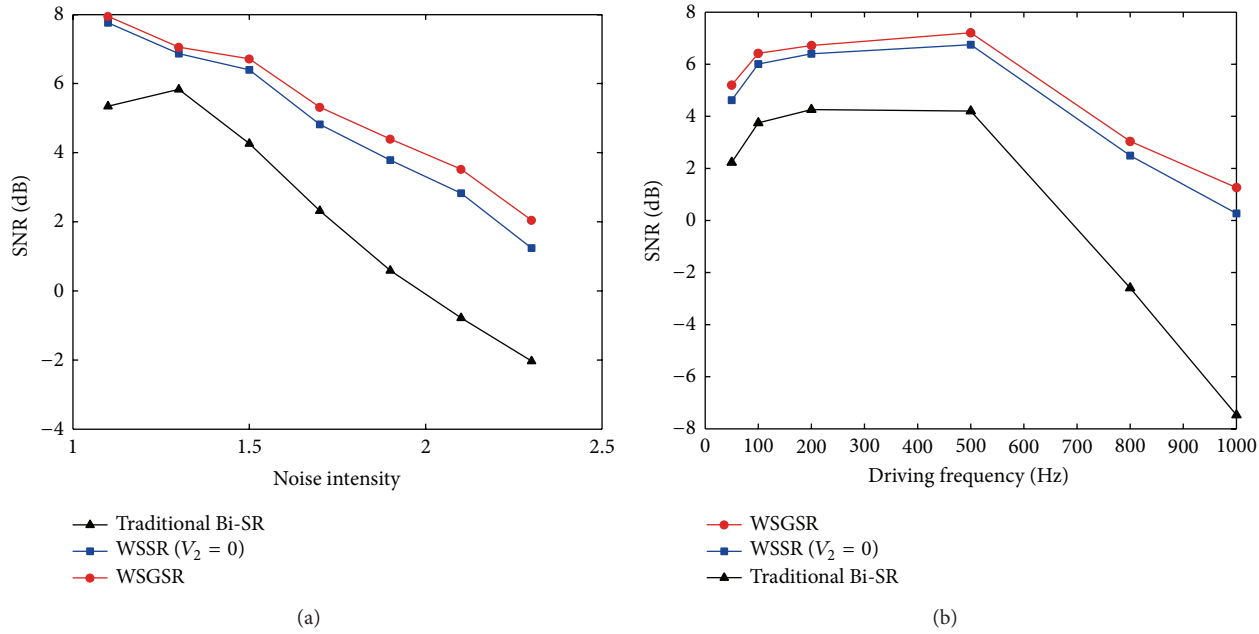
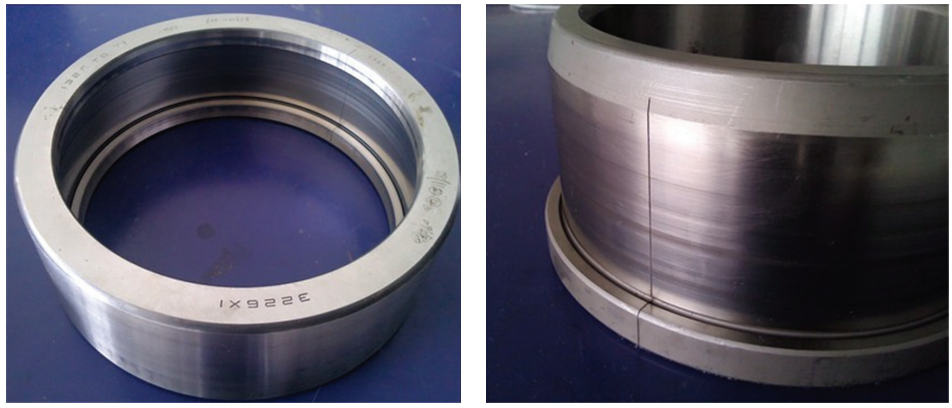


Figure 10: Output SNRs under different noise and driving frequency levels: (a) output SNR with the changing of noise intensity and (b) output SNR with the changing of driving frequency.

The train bearings used in this experiment are single row radial short cylindrical roller bearings, with the type of NJ(P)3226X1 with its detailed specification in Table 1. Firstly, we fabricate the artificial cracks of single line faults on the train bearing using wire-electrode cutting machine with the width of 0.18 mm on the outer and inner race, respectively, as shown in Figure 11. Then the faulty bearing is placed on an experiment platform in Figure 12. In the experiment, the acoustic bearing fault signal was acquired continuously by a microphone (type: 4944-A, manufactory: B&K Company) being mounted beside the outer race of the bearing placed at the radial direction with the distance of 150 mm. Except for the microphone, a National Instruments (NI) system (PXI-1033) equipped with PXI-4472 data acquisition card was used. According to Figure 12, the bearing could be applied with the mechanical load whose magnitude was measured by the pressure sensor and was displayed on a digital tube and it is driven by a motor, which was controlled by a frequency converter to match the anticipant rotating speed.

With all the introduced equipment, we can get the expecting bearing fault signal. In the experiment, the rotating speed is set at 1430 r/min, with a load of about 3 t and the sampling rate of 50 kHz. Through the calculation of bearing fault frequency equation with the parameters in Table 1, the value of the failure frequency with the defect on the outer race is 138.74 Hz and the value of the inner-race defective frequency is 194.93 Hz. Then after trying the experiment, we acquire the acoustic bearing signal with outer-race or inner-race defect. They are used to test the proposed WSGSR model and in order to have a comparison, the envelope signal and traditional bistable SR are also utilized to process the signal. The analysis results are described as below.

5.1. Results with Outer-Race Fault. The analyzed results of the outer-race defective train bearing signal using different methods are displayed in Figure 13. Figure 13(a) shows us the original signal and its spectrum. As a modulated signal, the structure resonance band of the raw signal is between 500 Hz and 2500 Hz where the main energy locates. But the fault signal cannot be found in both the waveform and the corresponding power spectrum. Making the signal band-pass filtered with band of [500 Hz, 2500 Hz] and demodulated we get the envelope signal with spectrum in Figure 13(b). Surrounded by the conspicuous noise components in both high and low frequency regions, the structural defective frequency f_{BPFO} can hardly be distinguished at the power spectrum with a SNR of −2.73 dB. Send the envelope signal into a traditional bistable SR system with parameters of a = 1.2, b = 46 which are optimized for the highest output SNR. The output in Figure 13(c) tells us that the noise components have been suppressed and the f_{BPFO} has been highlighted in the power spectrum with SNR of 1.58 dB, which implies that part of noise energies has been successfully transferred to the driving frequency. But the results are not satisfiable as there are still too many noises of low frequency. Subsequently, the same signal is analyzed by employing the proposed WSGSR method, and the results are exhibited in Figures 13(d) and 13(e). The first one is the output with V_2 = 0 which means a WSSR model with the other parameters optimized V_1 = 16, R_1 = 0.35, and a = 0.2. In this case the SNR gets further amplified to 2.98 dB. The second one is the optimal results of proposed method with parameters of V_1 = 2, R_1 = 0.05, a = 0.2, V_2 = 0.8, and R_2 = 0.25. In frequency domain the component of fault frequency has been successfully extracted with SNR of 5.11 dB. We can find a relatively



(a) (b)

Figure 11: Artificial cracks on the train bearing: (a) outer-race crack and (b) inner-race crack.

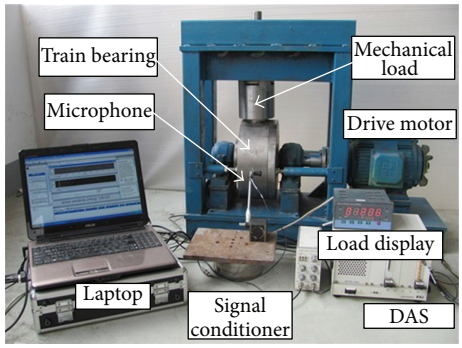


Figure 12: Experiment platform for bearing fault signal acquisition.

distinct f_{BPFO} in the spectrum implying that the proposed method is insensitive to the noise and could instead utilize the background noise to enhance the driving frequency. The differences between results in Figures 13(d) and 13(e) also give evidence of the amelioration compared with WSSR system.

5.2. Results with Inner-Race Fault. The analyzed results of the inner-race defective train bearing signal dealing with different systems are displayed in Figure 14. The original signal and its spectrum are provided in Figure 14(a). With a main energy band from 500 Hz to 2500 Hz, it delivers no useful information about the fault frequency in both time and frequency domain. Next the envelope signal and its corresponding power spectrum are displayed in Figure 14(b) filtered with band of [500 Hz and 2500 Hz]. In the spectrum, we can find a distinct low frequency signed f_r with a fairly high energy in the figure which indicates the rotating frequency of 23.8 Hz (1430 r/min). With interference of too much bootless frequencies as well as noise, we cannot figure out the fault frequency f_{BPFI} with SNR of −3.01 dB successfully from the envelope spectrum and the defect-induced impulses in time domain are not significant. In the results of traditional bistable SR with $a = 0.8$ and $b = 38$ which are optimized for the highest output SNR as in Figure 14(c), it gives a better result with SNR = −0.88 dB. But the interferences are still frustrating as lower frequencies which can still clearly be seen. The results of WSGSR show us a better situation. With fixing $V_2 = 0$ when the model acts as WSSR, Figure 14(d) provides a SNR of 2.92 dB with optimal parameters of $V_1 = 10$, $R_1 = 0.5$, and $a = 0.1$. But when we add it to 1 with the other parameters $V_1 = 12$, $R_1 = 1$, $a = 0.2$, and $R_2 = 0.45$ the system gets an optimal condition and the output becomes the one in Figure 14(e). It illustrates in the power spectrum that the f_{BPFI} has been amplified with a higher power than any other noise components with SNR of 4.76 dB and we can see the defect-induced impulses in time domain. The results show us the advantages of WSGSR model in bearing fault diagnosis embedded with heavy noises once again.

5.3. Discussion of Experimental Results. With the above subsections, the effectiveness and efficiency of the proposed weak signal detection strategy based on WSGSR have been verified by both the simulated signal and the practical bearing fault signals. The results show a similar verdict.

(1) Considering the two expressions of different SR models as in (6) and (7), it is obvious that the parameters a and b in traditional bistable SR are coupling which means they always work on the potential shape together. With all the parameters affecting potential model's features independently, the coupling effect which the parameters have in a traditional bistable SR model is eliminated well. For example, we can just tune V_2 to have different barrier height, R_1 to gain different well width, and a smaller a to make the well wall steeper. As a limitation, the optimal bistable potential that can enhance the periodic weak signal may not be accurately designed with the coupling effect of parameters, so the signal enhancement of the bistable SR is limited. For the WSGP, all the features can be independently designed by their corresponding coefficients. As a result, the WSGP can be adjusted to be matched to different input signal better.

(2) Compared to the shape of traditional bistable potential, the WSGD has a flatter well bottom and steeper well wall. This might make the particle oscillate more easily with less resistance and higher rebound force

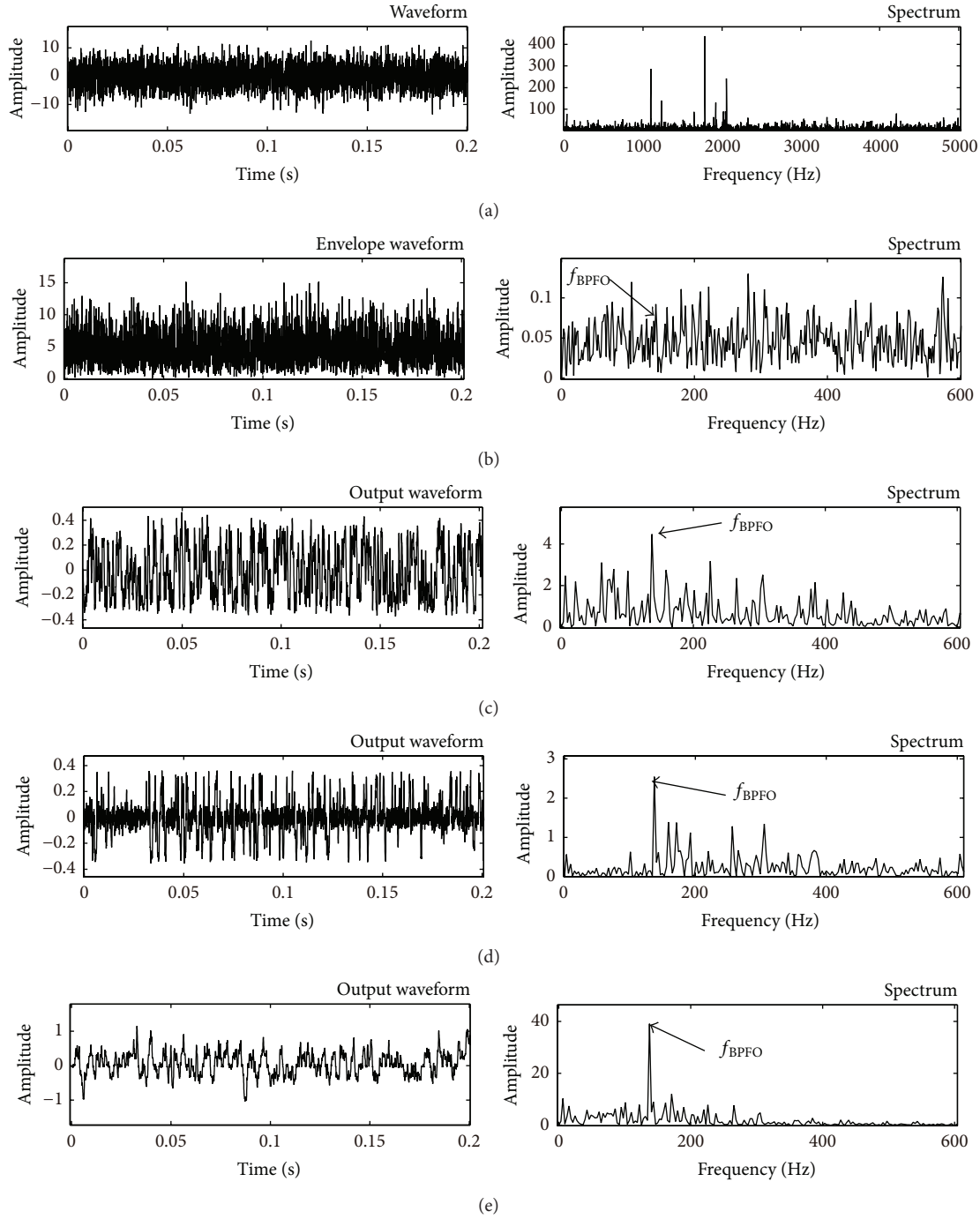


FIGURE 13: Optimal outputs of different models with outer-race fault signal: (a) original waveform and its power spectrum, (b) envelope signal and its spectrum filtered with [500 Hz and 2500 Hz], (c) output of bistable SR model and its spectrum with $a = 1.2$ and $b = 46$, (d) output of WSGSR model and its power spectrum with $V_1 = 16$, $R_1 = 0.35$, $a = 0.2$, and $V_2 = 0$ (WSSR), and (e) output of WSGSR model and its power spectrum with $V_1 = 2$, $R_1 = 0.05$, $a = 0.2$, $V_2 = 0.8$, and $R_2 = 0.25$.

which we have described before. The outputs verify our conjecture that the SR effect comes up more easily in the proposed model. The simulating and engineering results all provide the evidence that WSGSR results in a higher output SNR. More commendable, it can work with a wider range of noise intensity and driving frequency. In other words, even with heavier noise and higher frequency, the advantages of proposed method are more obvious and it can offer a better performance than the traditional one. This will owe to the special shape of WSGP.

(3) Besides, as a more general type, the WSGSR not only has the advantages of WSSR which is a particular case but also possesses a better performance. With utilizing of a potential barrier which can be turned

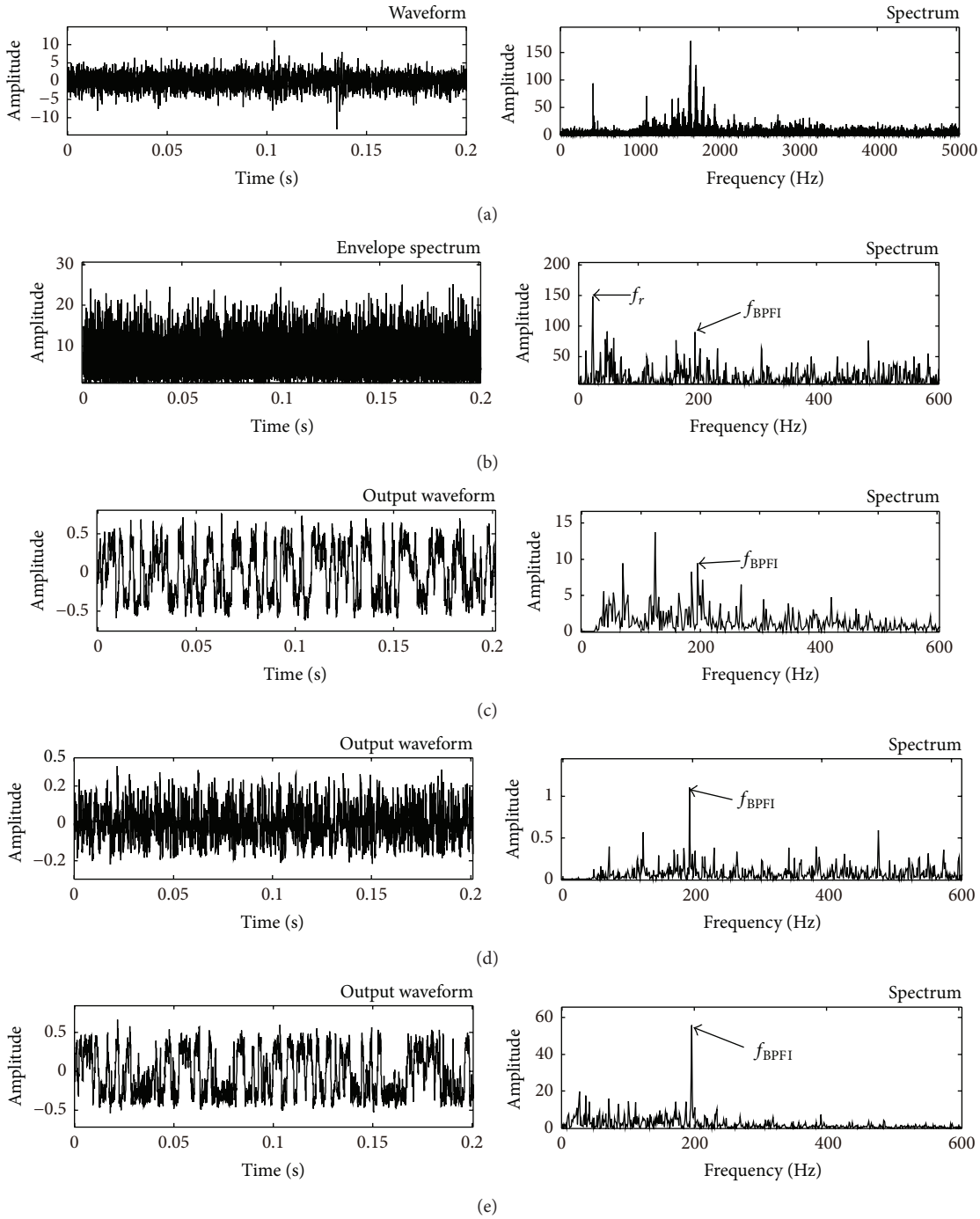


Figure 14: Optimal outputs of different models with inner-race fault signal: (a) original waveform and its power spectrum, (b) envelope signal and its spectrum filtered with [500 Hz and 2500 Hz], (c) output of bistable SR model and its spectrum with = 0.8, b = 38, (d) output of WSGSR model and its power spectrum with $V_1 = 10$, $R_1 = 0.5$, $a = 0.1$, and $V_2 = 0$ (WSSR), and (e) output of WSGSR model and its power spectrum with $V_1 = 12$, $R_1 = 1$, $a = 0.2$, $V_2 = 1$, and $R_2 = 0.45$.

by V_2 and R_2, the proposed potential model can easily change between monostable (WSSR) and bistable and has different features. This makes the system more perfect and has a large scope to tune the parameters when optimizing the SR system. The results also show that with the working of potential barrier, output SNR gets higher sometimes especially with heavier noise. It has a stronger ability in filtrating the characteristic frequency ignoring the components of lower or higher frequency. This may owe to the barrier that can weaken the uncorrelated components which can make the particle oscillate more easily without it.

(4) We have mentioned that when considering the amplitude of output spectrum in Figures 9, 13, and 14, we can easily find that the amplitude of driving frequency in WSGSR is more than 10 times larger than that of the WSSR ($V_2 = 0$) even the traditional bistable SR which may be the result of the existence of potential barrier. The higher amplitude means the higher energy on the output signal. So the WSGSR not only keeps the advantages of WSSR but also offers high output energy on driving frequency as the traditional bistable SR model which might benefit to the area of energy harvest. What is more is that we improve the WSSR model to a bistable one where many theories of bistable SR can be applied to, for example, the Kramer's rate. The general type offers us the possibility to have further research on the specific role that the potential shape plays on the performance of SR output signal.

(5) In engineering applications, the key is how to determine the model parameters. However, for an unknown system, we cannot obtain the fault samples in advance. In this case, we know in most of the practical applications that the fault information cannot be obtained easily before the diagnosis work. When processing the signal with band-pass filter and HT, we might not distinguish the fault frequency effectively due to heavy background noise. Then with just a SR model whose parameters are preset, we can extract the component which might not be so distinct. But with this information, we can use the component to adjust the parameters to the optimal combination according to the highest SNR which will make the fault frequency the clearest. If the component is exactly the periodic fault signal, it can be amplified to a high level by the optimal parameters but not vice versa. So the proposed method can offer us a more effective approach to make an accurate and reliable diagnosis.

However, the response of the WSGSR system is still complex and sometimes comes up with randomness as the acquired signal and noise always contain the uncertainty. Hence, a small variation of parameters may result in a quite different output like butterfly effect. What is more is that the model needs some mathematical analysis to support the certain effects that parameters' work has on the output. A deeper mechanism of the WSGSR from both the mathematical and the physical aspects should be investigated. These further studies will make the parameter selection and optimization more effective and gain a more satisfying output.

6. Conclusions

An improved potential model of WSGP is investigated to realize the SR effect instead of the traditional bistable one. The model of WSGSR with a particular form of WSSR ($V_2 = 0$) has been described in detail to explain its superior performance in bearing fault diagnosis. A diagnostic scheme is present with tuning the system's five parameters. Different from the traditional bistable SR, the features of potential model are affected by the parameters independently. For example, V_1 works on the well depth, R_1 works on the well width, a and R_2 affect the steepness of potential walls, and V_2 affects the barrier height. It means that an optimal state can be obtained by varying the parameters. During the searching process, a method to confirm the reference frequency to calculate SNR is put forward. The performance of the proposed method has been evaluated by both the simulated and the engineering bearing fault signals. The results among different models show a more efficient and continuous particle motion in WSGP benefiting from its distinct structure. Besides, the WSGSR is insensitive to the noise and owns the better capacity in detecting weak impulse signals and filtrating characteristic frequency with a higher output SNR under different noise and driving frequency levels. The potential barrier also makes it a better performance especially with heavier noise than WSSR. And hence, the driving frequency can be enhanced extremely while the noise interference can be suppressed. As a more general form, the proposed WSGSR offers a good performance in weak periodic signal amplification to diagnose bearing faults. It is reasonable to believe that this method can be used in mechanical diagnostic work especially when the acquired signal is seriously embedded with heavy background noises to amplify the weak periodic signal of fault or characteristic frequency.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China with Grant nos. 51075379 and 11274300. The authors also would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

[1] R. Benzi, A. Sutera, and A. Vulpiani, "The mechanism of stochastic resonance," *Journal of Physics A: Mathematical and General*, vol. 14, no. 11, pp. L453–L457, 1981.

[2] L. Gammaitoni, P. Hänggi, P. Jung, and F. Marchesoni, "Stochastic resonance," *Reviews of Modern Physics*, vol. 70, no. 1, pp. 223–287, 1998.

[3] Y. G. Leng, Y. S. Leng, T. Y. Wang, and Y. Guo, "Numerical analysis and engineering application of large parameter stochastic resonance," *Journal of Sound and Vibration*, vol. 292, no. 3–5, pp. 788–801, 2006.

[4] J. Tan, X. Chen, J. Wang et al., "Study of frequency-shifted and re-scaling stochastic resonance and its application to fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 23, no. 3, pp. 811–822, 2009.

[5] M. Lin and Y. M. Huang, "Modulation and demodulation for detecting weak periodic signal of stochastic resonance," *Acta Physica Sinica*, vol. 55, no. 7, pp. 3277–3282, 2006.

[6] J. Li, X. Chen, and Z. He, "Multi-stable stochastic resonance and its application research on mechanical fault diagnosis," *Journal of Sound and Vibration*, vol. 332, no. 22, pp. 5999–6015, 2013.

[7] Q. He, J. Wang, Y. Liu, D. Dai, and F. Kong, "Multiscale noise tuning of stochastic resonance for enhanced fault diagnosis in rotating machines," *Mechanical Systems and Signal Processing*, vol. 28, pp. 443–457, 2012.

[8] Q. He and J. Wang, "Effects of multiscale noise tuning on stochastic resonance for weak signal detection," *Digital Signal Processing*, vol. 22, no. 4, pp. 614–621, 2012.

[9] X. Zhang, N. Hu, L. Hu, and Z. Cheng, "Multi-scale bistable stochastic resonance array: a novel weak signal detection method and application in machine fault diagnosis," *Science China Technological Sciences*, vol. 56, no. 9, pp. 2115–2123, 2013.

[10] N. Tandon and A. Choudhury, "Review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings," *Tribology International*, vol. 32, no. 8, pp. 469–480, 1999.

[11] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems-Reviews, methodology and applications," *Mechanical Systems and Signal Processing*, vol. 42, no. 1-2, pp. 314–334, 2014.

[12] S. Lu, Q. He, F. Hu, and F. Kong, "Sequential multiscale noise tuning stochastic resonance for train bearing fault diagnosis in an embedded system," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, pp. 106–116, 2014.

[13] X. Zhang, N. Hu, Z. Cheng, and L. Hu, "Enhanced detection of rolling element bearing fault based on stochastic resonance," *Chinese Journal of Mechanical Engineering*, vol. 25, no. 6, pp. 1287–1297, 2012.

[14] L. Qiang, W. Taiyong, L. Yonggang, W. Wei, and W. Guofeng, "Engineering signal processing based on adaptive step-changed stochastic resonance," *Mechanical Systems and Signal Processing*, vol. 21, no. 5, pp. 2267–2279, 2007.

[15] H. L. He, T. Y. Wang, Y. G. Leng, Y. Zhang, and Q. Li, "Study on non-linear filter characteristic and engineering application of cascaded bistable stochastic resonance system," *Mechanical Systems and Signal Processing*, vol. 21, no. 7, pp. 2740–2749, 2007.

[16] S. Lu, Q. He, and F. Kong, "Stochastic resonance with Woods-Saxon potential for rolling element bearing fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 45, pp. 488–503, 2014.

[17] J. I. Deza, R. R. Deza, and H. S. Wio, "Wide-spectrum energy harvesting out of colored Lévy-like fluctuations, by monostable piezoelectric transducers," *EPL*, vol. 100, no. 3, Article ID 38001, 2012.

[18] M. J. Roberts, "Energy moments and cross sections for a gaussian potential," *Proceedings of the Physical Society*, vol. 82, no. 4, article 319, pp. 594–604, 1963.

[19] J. Adamowski, M. Sobkowicz, B. Szafran, and S. Bednarek, "Electron pair in a Gaussian confining potential," *Physical Review B*, vol. 62, pp. 13233–13233, 2000.

[20] W. Xie, "Two interacting electrons in a Gaussian confining potential quantum dot," *Solid State Communications*, vol. 127, no. 5, pp. 401–405, 2003.

[21] G. Stephenson, "Eigenvalues of the Schrödinger equation with a Gaussian potential," *Journal of Physics A: Mathematical and General*, vol. 10, no. 12, pp. L229–L232, 1977.

[22] M. Cohen, "On the Schrödinger equation with a Gaussian potential," *Journal of Physics A: Mathematical and General*, vol. 17, no. 3, pp. L101–L104, 1984.

[23] R. D. Woods and D. S. Saxon, "Diffuse surface optical model for nucleon-nuclei scattering," *Physical Review*, vol. 95, no. 2, pp. 577–578, 1954.

[24] A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical Systems and Signal Processing*, vol. 20, no. 7, pp. 1483–1510, 2006.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 329458, 11 pages
http://dx.doi.org/10.1155/2014/329458

Research Article

Fault Detection Enhancement in Rolling Element Bearings via Peak-Based Multiscale Decomposition and Envelope Demodulation

Hua-Qing Wang,[1] Wei Hou,[1] Gang Tang,[1] Hong-Fang Yuan,[2] Qing-Liang Zhao,[1] and Xi Cao[2]

[1] *School of Mechanical and Electrical Engineering, Beijing University of Chemical Technology, Beijing 100029, China*
[2] *School of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China*

Correspondence should be addressed to Gang Tang; tanggang@mail.buct.edu.cn

Received 31 March 2014; Accepted 7 May 2014; Published 27 May 2014

Academic Editor: Ruqiang Yan

Vibration signals of rolling element bearings faults are usually immersed in background noise, which makes it difficult to detect the faults. Wavelet-based methods being used commonly can reduce some types of noise, but there is still plenty of room for improvement due to the insufficient sparseness of vibration signals in wavelet domain. In this work, in order to eliminate noise and enhance the weak fault detection, a new kind of peak-based approach combined with multiscale decomposition and envelope demodulation is developed. First, to preserve effective middle-low frequency signals while making high frequency noise more significant, a peak-based piecewise recombination is utilized to convert middle frequency components into low frequency ones. The newly generated signal becomes so smoother that it will have a sparser representation in wavelet domain. Then a noise threshold is applied after wavelet multiscale decomposition, followed by inverse wavelet transform and backward peak-based piecewise transform. Finally, the amplitude of fault characteristic frequency is enhanced by means of envelope demodulation. The effectiveness of the proposed method is validated by rolling bearings faults experiments. Compared with traditional wavelet-based analysis, experimental results show that fault features can be enhanced significantly and detected easily by the proposed method.

1. Introduction

A rolling bearing is one of the most widely used elements in rotating machinery. As a critical component, it carries most of the load during the running of rotating machinery. If rolling bearing fails, serious problems arise, which will in turn result in the decrease of production efficiency and large economic loss. Records show that faulty bearings contribute to about thirty percent of failures in rotating machinery. Thus, it is of great importance to study the effective fault diagnosis approaches for rolling bearings.

Various methods have been developed for bearing fault diagnosis and condition monitoring, such as vibration monitoring, temperature monitoring, chemical analysis, acoustic emission monitoring, sound pressure monitoring, and laser monitoring [1]. Vibration signal analysis is one of the most efficient methods thanks to the useful information of machine's work status carried by vibration signals [2–4]. To extract fault information, for example, partial defects of bearings, vibration signals are usually processed in time domain, frequency domain, or both [5, 6]. In time domain, some statistical parameters, for example, probability density, kurtosis, root mean square, or skewness, have been introduced into bearing defect detection [7]. However, these fault indicators are not effective in all cases, especially for weak fault signals. Therefore, analysis approaches in frequency domain are developed to detect bearing defect fault. Based on the efficiency of modern fast Fourier transform (FFT), spectral analysis of vibration signals is widely used to extract characteristic defect frequencies. In fact, bearing fault signal is

amplitude modulated at its characteristic defect frequencies, so a preprocessing of demodulation should be performed before FFT applied. Traditionally, envelope demodulation is used to detect the fault frequency [5, 8]. For obvious fault features, traditional envelope demodulation has notable advantages in fault type recognition. However, it does not work well in many cases for the detection of weak fault features. To solve this problem, various approaches have been developed to denoise vibration signals and to enhance the characteristic defect frequency. Cyclostationary analysis [9], for example, spectral autocorrelation analysis [10], is such a way. In recent years, some methods based on time-frequency analysis, such as empirical mode decomposition (EMD) [11–13], local mean decomposition (LMD) [14, 15], and wavelet transform [16, 17], are also developed to enhance fault features and detect failures. Both EMD and LMD are sensitive to noise level, but they are so hard to implement properly since these algorithms contain imperfections. In addition to the mentioned signal processing approaches, some other strategies are also introduced for fault features extraction as well as pattern recognition, for example, symptom parameter waves [18] and fuzzy diagnosis method [19].

As mentioned above, signal's decomposition or sparse representation is the basis of those methods. However, because of the existing high frequency components, vibration signals are not always sparse enough after decomposition. As a popular strategy for bearing vibration signals' processing, wavelet-based analysis is effective for feature extraction from smooth signals, signal denoising, and fault feature enhancement [20–22]. But it cannot always represent vibration signals sparsely enough, especially for those mixed by plenty of interferential signals generated by other related rolling elements. This will be illustrated further with a typical example in the next section. In related research areas, a so-called peak transform based on piecewise curve recombination has been developed. The idea is to adjust piecewise curves to improve the smoothness of the target signals or images, and it has been used in 2D image representation and coding [23, 24]. To our knowledge, there is no report in the literature by far on its applications to bearing vibration signal analysis.

In order to overcome those disadvantages of wavelet-based methods for weak signals processing of bearing faults, this paper developed a new multiscale decomposition strategy with sparsity-promoting by recombination of piecewise signals under different frequencies. The new strategy improves the sparsity of original vibration signals and is of great significance in fault signals denoising and enhancement. Then the wavelet coefficients of vibration signals will be sparse enough, and a threshold method based on wavelet decomposition can eliminate most noise. Finally, the FFT-based Hilbert transform is conducted for signal demodulation and extracting fault signature of characteristic defect frequency. Fault features will be significantly enhanced and easily detected through the proposed method.

The paper is organized as follows. Section 2 illustrates the sparseness of vibration signals in wavelet domain, which is also the motivation of this work. The proposed method of peak-based multiscale decomposition and envelope demodulation is introduced in Section 3. Some experimental results to show the effectiveness of the method are presented in Section 4. Finally, discussions and conclusions are presented in Section 5.

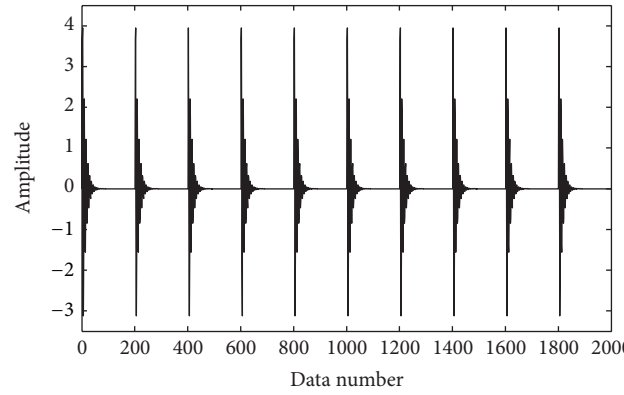


Figure 1: Vibration waveform of a bearing with a point defect on the outer race.

2. Vibration Signal of Rolling Bearing Faults and Its Sparsity in Wavelet Domain

Vibration signal is always acquired by sensors related to acceleration, velocity, or displacement measurement. A series of signal processing steps are then applied to make the fault features easy to be distinguished. Basically, the key of vibration signal analysis is to find out appropriate parameters describing the running conditions of bearings clearly. The commonly used characteristic defect frequency is one of the most effective indicators for rolling bearing's fault. If local defects exist in a bearing, the measured vibration signal will be amplitude modulated. Its carrier wave is a kind of the bearing's inherent vibration with high frequency, while its modulating wave is a pass vibration corresponding to the local defects, whose frequency is called ball pass frequency or characteristic defect frequency. The latter is usually utilized to determine whether a fault exists, and it is also adopted by this paper.

In order to reduce the potential impact of signal attenuation through complex paths, sensors are often placed as close to the bearings as possible, for example, bearing housing. However, transmission path still exists between the impact point and the vibration measurement point, which would make signals more complex and hard to be demodulated. If defects exist, a fault in one surface of bearings would strike another; then a force impulse will be generated and flowing resonances in the bearing and machine will also be excited [25]. A series of impulse responses will be generated and last continuously, which may be amplitude modulated caused by the fault passing through the load zone or the mentioned varying transmission paths. Therefore, the ideal bearing defect signal is a kind of periodical impulse-like signal with high frequency nonsmooth components. A typical vibration signal of a bearing is shown in Figure 1; the vibration waveform is obviously impacted by an existing point defect on the outer race. In practice, in addition to bearing condition

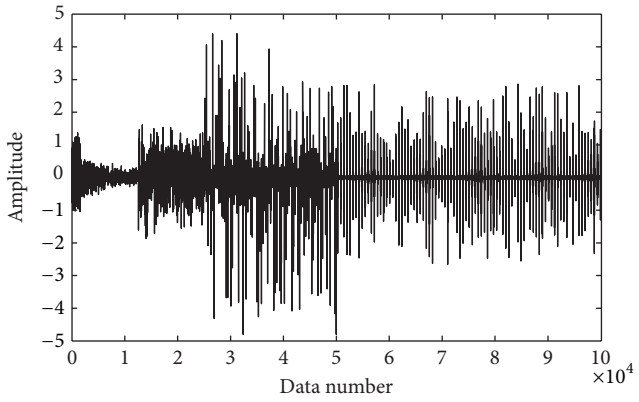


Figure 2: Wavelet coefficients of the bearing fault signal in Figure 1.

information, the measured vibration signals are often mixed with lots of interfering information from other components. Meanwhile, various types of noise also make the signal more complicated. In addition to the noise introduced by measurement instruments, interfering signals can also be generated accompanied with the existence of transmission path between local impact point and vibration measurement point. In a word, bearing vibration signal is nonstationary with wide frequency band, which makes the bearing fault signature extremely weak to be detected.

Existing wavelet-based denoising methods mainly rely on the sparseness of wavelet coefficients [17]. If a signal is not sparse enough in wavelet domain, it becomes challenging for feature extraction by wavelet analysis due to the difficulties to distinguish required features from large number of wavelet coefficients. Common vibration signals from faulty bearings are often composed of impulse components. A typical vibration signal is shown in Figure 1, and its wavelet coefficients are shown in Figure 2, from which we can see that most coefficients are nonzero, that is, not sparse enough. Experiments show that this is applicable to fault vibration signals of inner race, outer race, and ball element. Thus it is important to find out a proper way to improve the sparseness, as well as reducing noise and enhancing the weak fault features.

3. The Proposed Method

3.1. The Basic Principle of Peak-Based Piecewise Recombination. Peak-based piecewise recombination (PPR) or so-called peak transform is a kind of nonlinear geometric transform to make the signal smoother through transforming high frequency subsections to low frequency ones. Here, a brief description of PPR is given as follows; for more information please refer to [23].

For a continuous function $f(x)$ that has N-point peaks (or break points), x_i $(i = 1 \cdots N)$ is defined over $[a, b]$. In fact, $f(x)$ can be viewed as a piecewise function with $N + 1$ combined curves, where the N peaks' positions are the connection points. Then the ith piecewise curve can be called subfunction $f_i(x)$ over interval $[a, b]$. The N-point forward

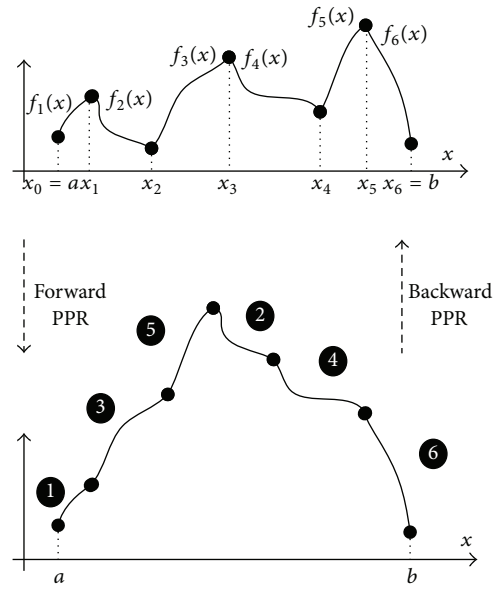


Figure 3: Schematic diagram of five-point peak-based piecewise recombination.

transform of peak-based piecewise recombination can be defined as

$$PF\left[f\left(x\right)\right] = g_o\left(x\right) \oplus g_e\left(x\right), \tag{1}$$

where

$$\begin{aligned} g_o\left(x\right) &= f_1\left(x\right) \oplus f_3\left(x\right) \oplus \cdots \oplus f_{2\lfloor (N-1)/2 \rfloor + 1}\left(x\right), \\ g_e\left(x\right) &= f_2\left(x\right) \oplus f_4\left(x\right) \oplus \cdots \oplus f_{2\lfloor N/2 \rfloor}\left(x\right) \end{aligned} \tag{2}$$

are the cascades of all odd- and even-numbered curve segments, respectively. Here, $2\lfloor (N-1)/2 \rfloor + 1$ and $2\lfloor N/2 \rfloor$ are, respectively, the largest odd and even integers that are less than or equal to N.

Geometrically speaking, as illustrated in Figure 3, all piecewise curves with similar positive slopes are grouped together and cascaded to be a new curve, while all negative ones are grouped into another new curve; finally, these two new subcurves are recombined to be a new curve. It is apparent that the recombination only changes the order of the piecewise curves and is reversible. The backward transform can be done by simply recascading them according to their original orders, which is symbolized as $PF^{-1}[f(x)]$ according to (1).

It should be noted that, although the forward and backward transforms discussed above are defined for curve functions, they can also be defined in similar manners for discrete-time signals of rolling bearings' vibration.

3.2. Wavelet Decomposition. Wavelet $\psi_{(a,b)}(t)$ is obtained by translation and dilation based on a defined single function $\psi(t)$ as

$$\psi_{(a,b)}\left(t\right) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right), \tag{3}$$

where $a > 0$ is the so-called scaling parameter and $b \in R$ is the parameter denoting time localization, which can be continuous or discrete. $\psi(t)$ is called "mother wavelet" to generate wavelets $\psi_{(a,b)}(t)$.

Given a signal $x(t)$ with finite energy, the wavelet transform with analytic wavelet $\psi(t)$ can be viewed as the convolution of $x(t)$ with a scaled and conjugated wavelet

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\, \psi^* \left(\frac{t-b}{a} \right) dt, \tag{4}$$

where $\psi^*(t)$ stands for the complex conjugation of $\psi(t)$.

The wavelet transform $W(a,b)$ can be considered as a function of translation b at each scale a. From (4), we can see that wavelet transform is a kind of time-frequency analysis or a time-scaled analysis. Different from analysis either in time domain or in frequency domain, or even different from that in time-frequency domain but with fixed-length windows by short time Fourier transform (SFFT), vibration signals can be decomposed by wavelet transform with multiscale analysis through dilation and translation, so that the time-frequency features of vibration signals can be more effectively extracted.

Wavelet transform is also reversible as follows:

$$x(t) = C_\psi^{-1} \iint W(a,b)\, \psi_{(a,b)}(t)\, \frac{da}{a^2} db, \tag{5}$$

where

$$C_\psi = \int_{-\infty}^{\infty} \frac{\left| \widehat{\psi(\omega)} \right|^2}{|\omega|}\, d\omega < \infty,$$
$$\widehat{\psi}(w) = \int \psi(t) \exp(-jwt)\, dt \tag{6}$$

which provides the possibility to reconstruct the original vibration signals.

3.3. Envelope Demodulation. As mentioned in the previous sections, if bearing defects exist, the measured vibration signal would be amplitude modulated at its characteristic defect frequency. The modulating wave is a pass vibration signal corresponding to local defects, and various demodulation techniques have been developed for the separation. In this work, an envelope demodulation method based on Hilbert transform is employed.

For a continuous time signal $x(t)$, the Hilbert transform $\widehat{x(t)}$ is defined as

$$\widehat{x(t)} = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{x(\tau)}{(t-\tau)}\, d\tau. \tag{7}$$

Then, combine $x(t)$ and $\widehat{x(t)}$ to form a new analytic signal

$$g(t) = x(t) + j\widehat{x(t)}. \tag{8}$$

The envelope of $x(t)$ is defined as the amplitude of $g(t)$:

$$A(t) = \sqrt{[x(t)]^2 + [x(t)]^2}. \tag{9}$$

After Hilbert transform, envelope spectra could be obtained by FFT to extract characteristic defect frequencies.

3.4. Enhancement Frame with Peak-Based Multiscale Decomposition and Envelope Demodulation. Wavelet transform is a kind of time-frequency analysis method for signal denoising, by adjusting the discrete detail coefficients and approximation coefficients obtained from multilevel decomposition. This scheme is called threshold processing, one of whose premises is the sparse representation of signals in wavelet domain. However, different from that of perfect bearing, the vibration signal energy of bearings with defects increases at high frequencies. Thus, vibration signals are not always sparse in wavelet domain. Piecewise curve recombination can convert energies at high frequencies to ones at low frequencies. In this way, signal energies are focused in low frequency subbands, so that the piecewise recombination can improve signal's smoothness as well as its sparseness in wavelet domain.

The present paper proposed an enhancement frame for vibration signals by the combination of peak-based multiscale decomposition and envelope demodulation. The procedure is described as follows.

(1) Peak-based piecewise recombination of vibration signal: let $x(t)$ denote the measured vibration signal, firstly; then, transform the original signal $x(t)$ to smooth signal $PF[f(x)]$ through piecewise recombination according to Equation (1). The key is to determine the peaks and to recombine different segments. In the present work, a cascading strategy related to local minimum and maximum is employed. The endpoint of a segment is chosen as a peak. If two adjacent peaks are equal, one of them is defined as the peak.

(2) Wavelet decomposition of the new recombined signal: choose a wavelet basis (e.g., Morlet wavelet or Db4 wavelet); then decompose the new signal $PF[f(x)]$ at level N:

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} PF[x(t)]\, \psi^* \left(\frac{t-b}{a} \right) dt. \tag{10}$$

(3) Coefficients threshold: set a threshold value and apply it to the approximation coefficients at level N and the detail coefficients at levels 1 to N. The hard threshold can be expressed as

$$y_s = \begin{cases} y & |y| > t \\ 0 & |y| < t, \end{cases} \tag{11}$$

where t is the selected threshold, y is the original wavelet coefficients, and y_s is the wavelet coefficients after thresholding. In this work, the value which covers about 95% of the smallest wavelet coefficients is set as the threshold value. Then 5% of the largest wavelet coefficients are kept as nonzero ones to reserve the signal feature.

(4) Signal recovery: reconstruct the vibration signal based on the modified approximation coefficients at level N and the modified detail coefficients at levels 1 to N:

$$PF'[x(t)] = C_\psi^{-1} \iint W'(a,b)\, \psi_{(a,b)}(t)\, \frac{da}{a^2} db. \tag{12}$$

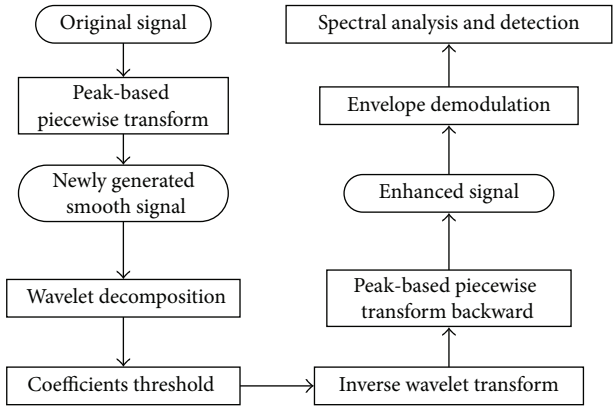


Figure 4: Flowchart of the proposed enhancement scheme for weak fault signal.

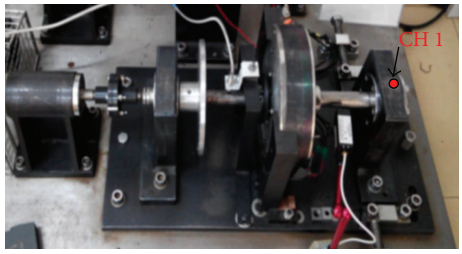


Figure 5: Fault rig of a roller element bearing.

(5) Piecewise signal backward transform: put the processed signal segments back to their original positions; then a new denoised vibration signal $x'(t)$ is generated.

(6) Envelope demodulation: according to (9), use Hilbert transform to obtain the envelope $A(t)$ of signal $x'(t)$.

(7) Extraction of characteristic defect frequency: get the envelope spectra via Fourier analysis and distinguish different fault types of rolling bearings.

In summary, these steps are shown as in Figure 4.

4. Experimental Results and Discussions

4.1. Experiment Setup. To verify the effectiveness of the proposed method, experiments are designed for fault rigs of roller element bearing, as shown in Figure 5. The rig is composed of a motor, a coupling, a rotor, and a shaft with two roller bearings. Basically, the sensors are located at the position near the bearings to mitigate the effects of signal attenuation. Usually, bearing housing is one of the best locations for bearing arrangement. Therefore, here, an accelerometer is located on the bearing housing to acquire the vibration signals.

In the experiments, a single point defect is introduced in inner raceway, outer raceway, and ball element of different bearings, using electron-discharge machining with fault widths of 3 mm, 7 mm, and 7 mm and depths of 5 mm, 25 mm, and 25 mm, respectively.

Vibration signals are measured by an accelerometer, located at the top of the bearing house (CH1, as shown in Figure 5). In all experiments, the sample frequency is 100 kHz and the shaft speed is finite, 1300 rpm. Four common conditions are studied in the present paper, including a perfect healthy bearing and three other bearings with point defect on the outer race, on the inner race bearing, and on the ball element.

In theory, elements of roller bearings have their own specific rotational frequencies, which may appear in envelope spectra if defects exist. These defect frequencies are called ball pass frequencies. For a bearing with a stationary outer race, these frequencies can be given by the following formulas:

$$\text{outer race defect frequency, } w_{\text{od}} = \frac{Zw_s}{2d}\left(1 - \frac{d}{D}\cos\alpha\right),$$

$$\text{inner race defect frequency, } w_{\text{id}} = \frac{Zw_s}{2}\left(1 + \frac{d}{D}\cos\alpha\right),$$

$$\text{rolling element defect frequency, } w_{\text{re}} = w_s\frac{D}{2d} \times\left(1 - \frac{d^2}{D^2}\cos^2\alpha\right), \tag{13}$$

where w_s is the shaft rotation frequency in rad/s, d is the diameter of rolling element, D is the pitch diameter, Z is the number of rolling elements, and α is the contact angle.

In the present experiments, outer race defect frequency, inner race defect frequency, and rolling element defect frequency are 86.32 Hz, 145.84 Hz, and 51.13 Hz, respectively.

4.2. Vibration Signals of Typical Bearing Faults and Their Envelop Spectra. The time-domain waveforms of bearing vibration signals obtained from accelerometers are shown in Figure 6, from which we can see that, different from that of perfect bearings as shown in Figure 6(a), obvious impulsive phenomenon exists in faulty vibration signals of fault bearings as shown in Figures 6(b), 6(c), and 6(d).

To investigate the differences of the signals and to distinguish different fault types, envelope demodulation is applied first to vibration signals of the three typical bearing faults, that is, outer race fault, inner race fault, and roller element fault. The results are shown in Figures 7, 8, and 9. In Figure 7, the frequency magnitude of outer race defect is lower than the second harmonic frequency of outer race defect. In Figure 8, although the inner race defect frequency slightly sticks out, it is still possibly confused by the sidelobes with almost similar magnitudes. In Figure 9, it is almost impossible to figure out the calculated ball pass frequency of 51.13 Hz. From the figures, we can see that, simply based on envelope demodulation, it is difficult to identify the fault frequencies from those of other components, especially for weak fault signals.

4.3. Wavelet Coefficients Distributions of the Vibration Signals with and without Peak-Based Piecewise Recombination. Before envelope demodulation, wavelet decomposition is

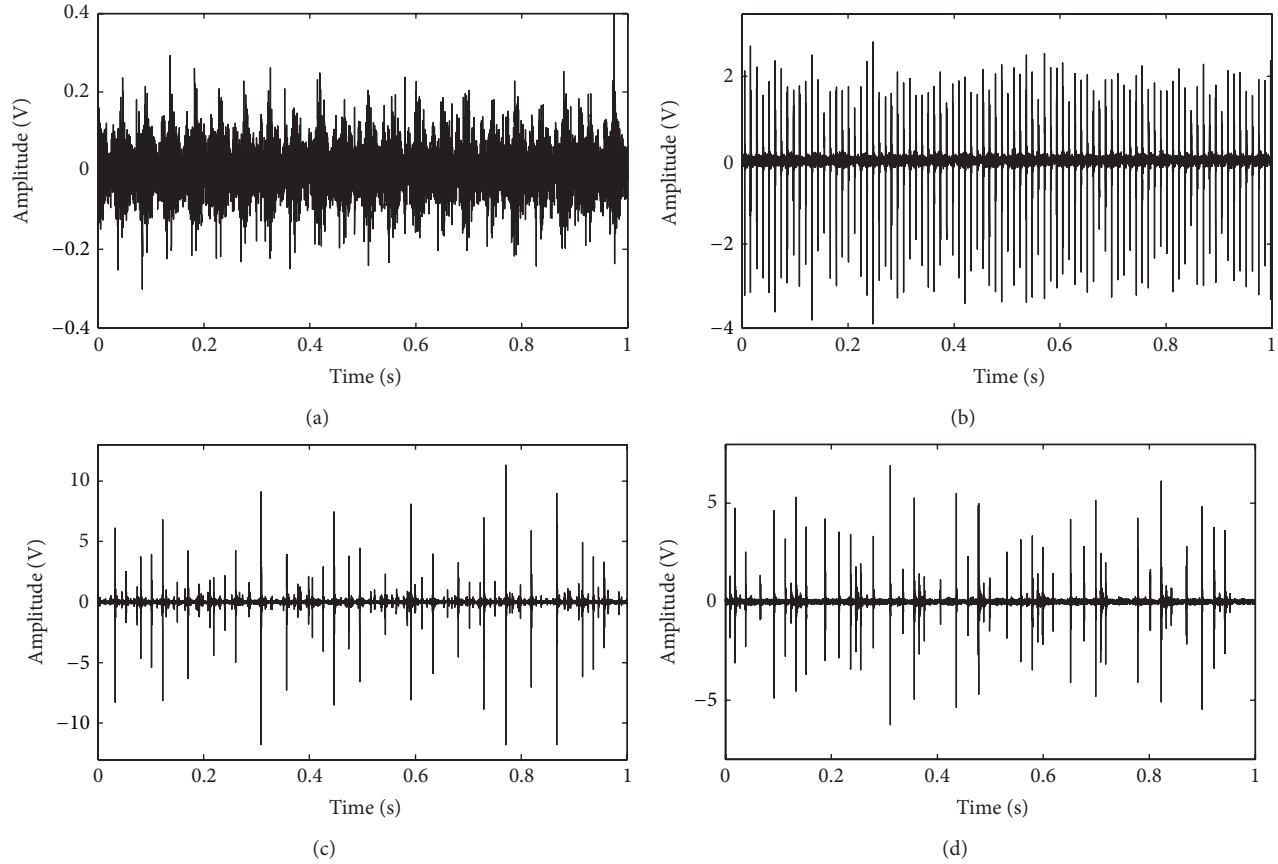


Figure 6: Vibration signals of rolling bearings: (a) a normal bearing and bearings with a point defect on (b) outer race, (c) inner race bearing, and (d) ball element.

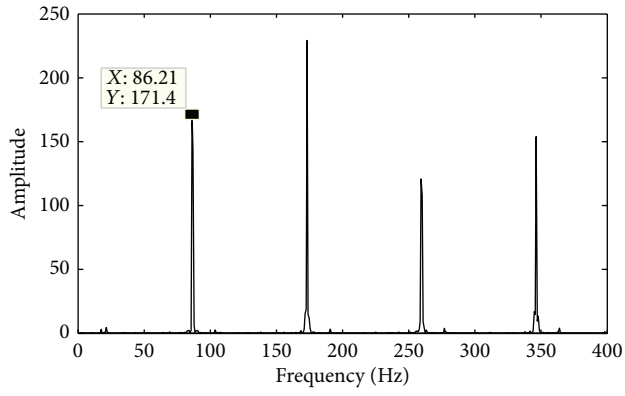


Figure 7: Envelope spectra of vibration signal in Figure 6(b) induced by outer race fault.

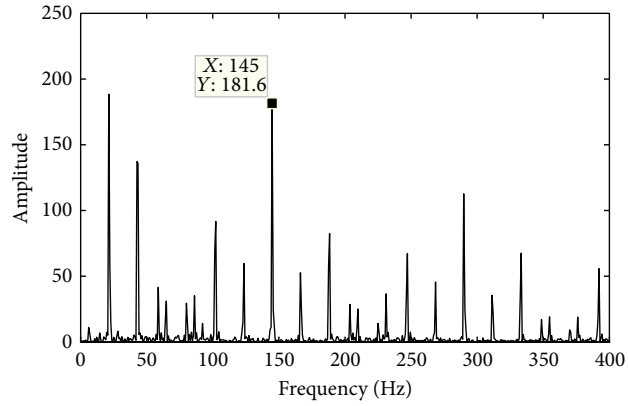


Figure 8: Envelope spectra of vibration signal in Figure 6(c) induced by inner race fault.

always applied to denoise and enhance the weak fault signals. However, vibration signals are not always sparse enough in wavelet domain, which would affect the effectiveness of the enhancement. As mentioned above, peak-based piecewise recombination can help to improve the smoothness of vibration signals and make their energies more concentrated in wavelet domain. Comparisons between wavelet decompositions with and without peak-based piecewise recombination are illustrated in this section.

Taking outer race fault as an example, the time-domain vibration signal of Figure 6(b) is converted to that of Figure 10 after applying piecewise recombination. The signal becomes much smoother than the original signal after piecewise

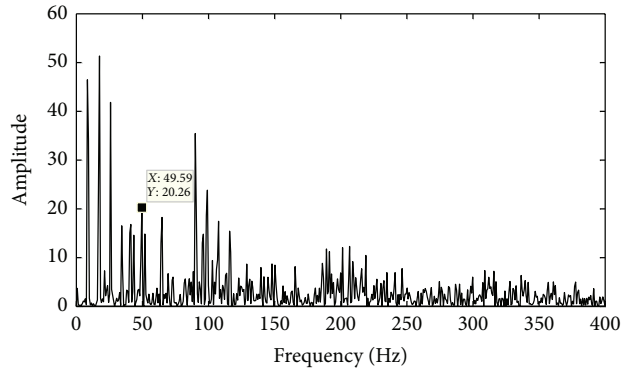


Figure 9: Envelope spectra of vibration signal in Figure 6(d) induced by roller element fault.

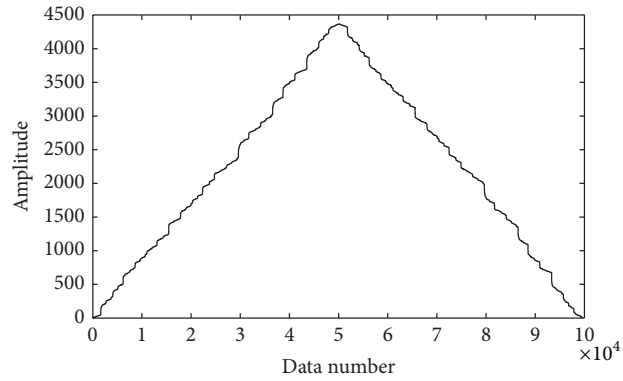


Figure 10: Vibration signal induced by outer race fault in Figure 6(b) after peak-based piecewise recombination.

Table 1: Energy distributions of the outer race fault signal's wavelet coefficients.

Frequency band	Low frequency	High frequency
Energies without PPR	49.31%	50.69%
Energies with PPR	99.9%	0.1%

recombination; that is, high frequency components have been converted to low frequency ones.

For comparisons, then wavelet transform is applied to signals in Figures 6(b) and 10, respectively, where 6-level decomposition with db4 wavelet basis is employed. Both distributions of the wavelet coefficients are shown in Figure 11. Without piecewise recombination, most wavelet coefficients are nonzero, that is, not sparse enough, as depicted in Figure 11(a). However, it is improved significantly with preprocessing of peak-based piecewise recombination, as depicted in Figure 11(b). Energies of the signal in wavelet domain are concentrated in only a few low frequency intervals, and almost all detail coefficients at scales 1 to 6 are close to zero. It implies that peak-based piecewise recombination can significantly improve the sparseness of vibration signals in wavelet domain.

Statistically, energy distributions of wavelet coefficients obtained by the two methods are listed in Table 1, from

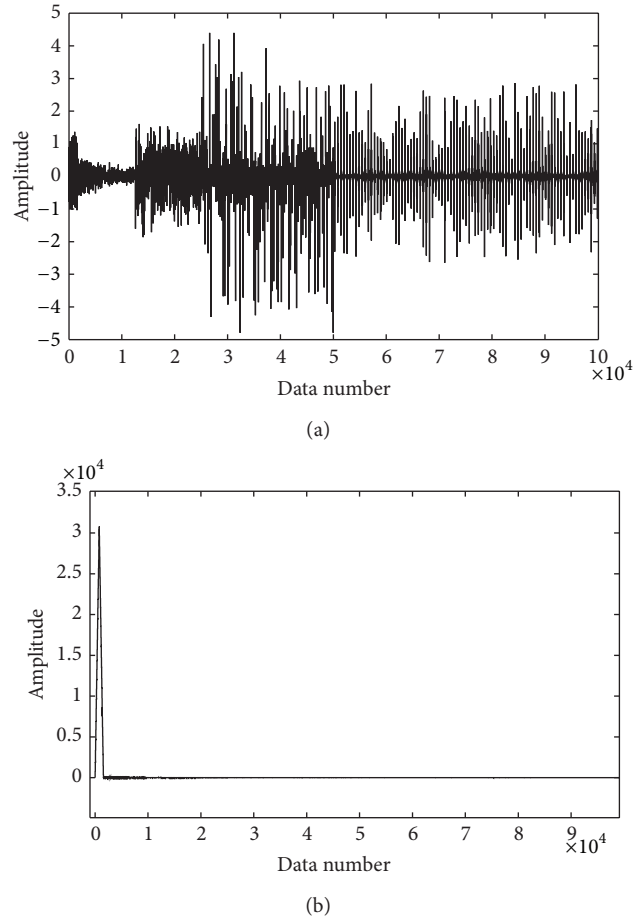


Figure 11: Wavelet coefficients of the vibration signal induced by outer race fault: (a) without PPR preprocessing and (b) with PPR preprocessing.

which we can see that, after employing peak-based piecewise recombination (PPR) preprocessing, energies of the signal are concentrated into only a few coefficients in wavelet domain, 99.9% in low frequency bands, compared with 49.31% without peak-based piecewise recombination.

The experimental results of inner race fault and roller element fault are similar to those of the outer race fault, which are shown in Figures 12 and 13.

4.4. Comparisons between Traditional Wavelet-Based Detection Method and the Proposed Strategy with Envelop Demodulation. Envelop demodulation is always combined with wavelet-based denoising and enhancement methods to analyze vibration signals. Comparisons of the results after envelop demodulation based on the results of Section 4.3 are presented in this section.

Envelop demodulation is applied to the signals reconstructed from the thresholding of wavelet coefficients, and a backward peak-based piecewise recombination is implemented for the proposed strategy. The demodulated envelope spectra are shown in Figures 14–16. In Figure 14(a), although the fault frequency can be figured out, the magnitude of

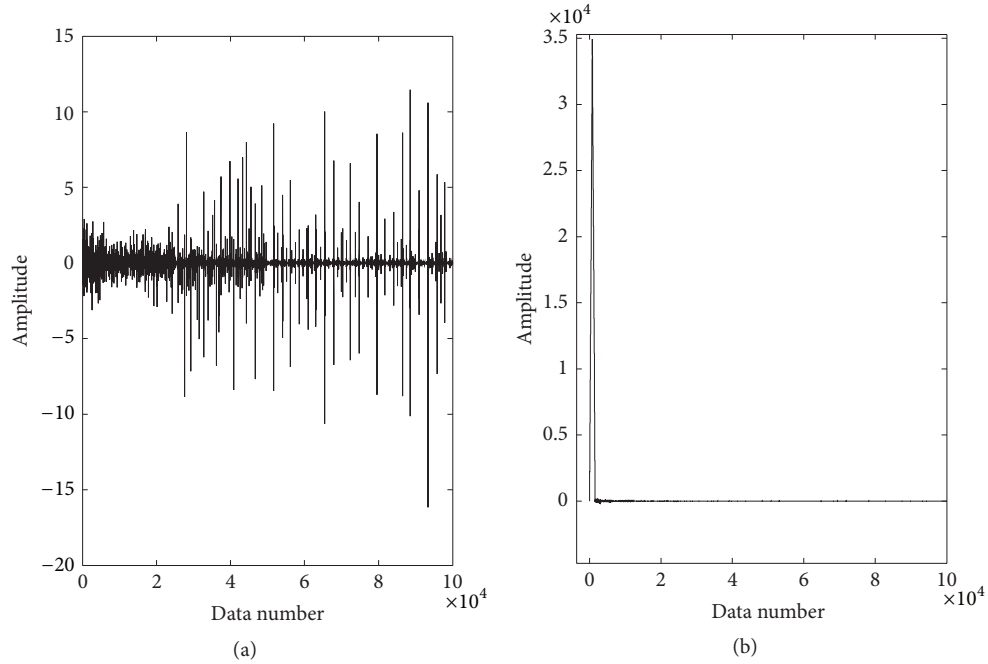


Figure 12: Wavelet coefficients of the vibration signal induced by inner race fault: (a) without PPR preprocessing and (b) with PPR preprocessing.

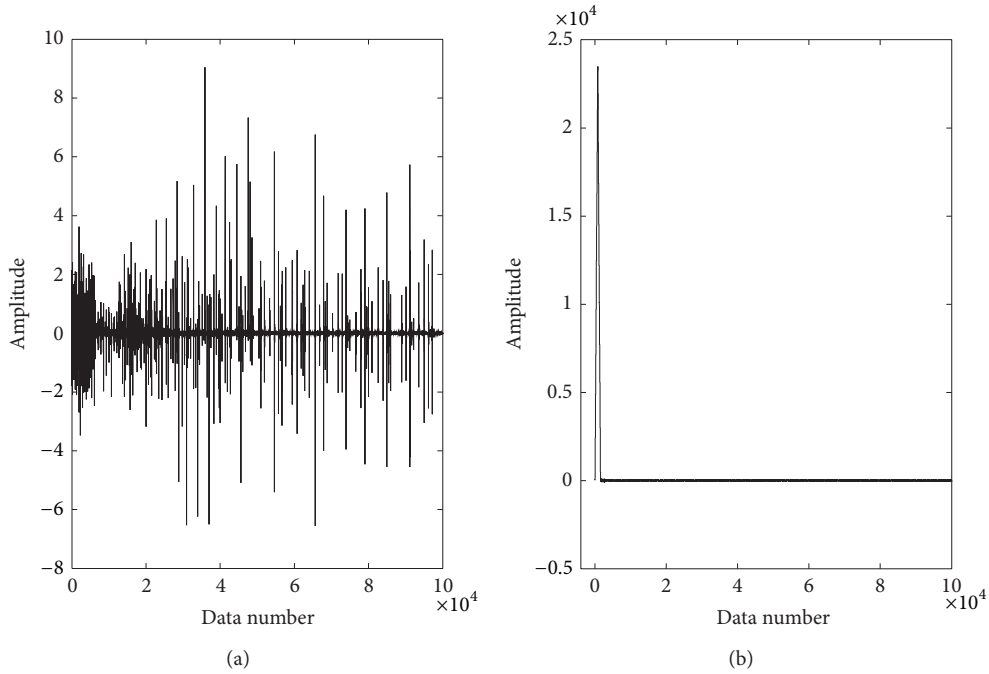


Figure 13: Wavelet coefficients of the vibration signal induced by roller element fault: (a) without PPR preprocessing and (b) with PPR preprocessing.

the outer race defect frequency is not the most remarkable one among all frequency components; thus, it is not always certain that a bearing defect exists in outer race. With the proposed method presented in this paper, the magnitude of outer race defect frequency becomes more prominent, where the magnitudes of outer race defect frequency have been enhanced and are more easily detected, as Figure 14(b) depicted. The proposed strategy is also more effective for inner race and roller bearing fault detection of rolling bearings, and the results are shown as in Figures 15 and 16.

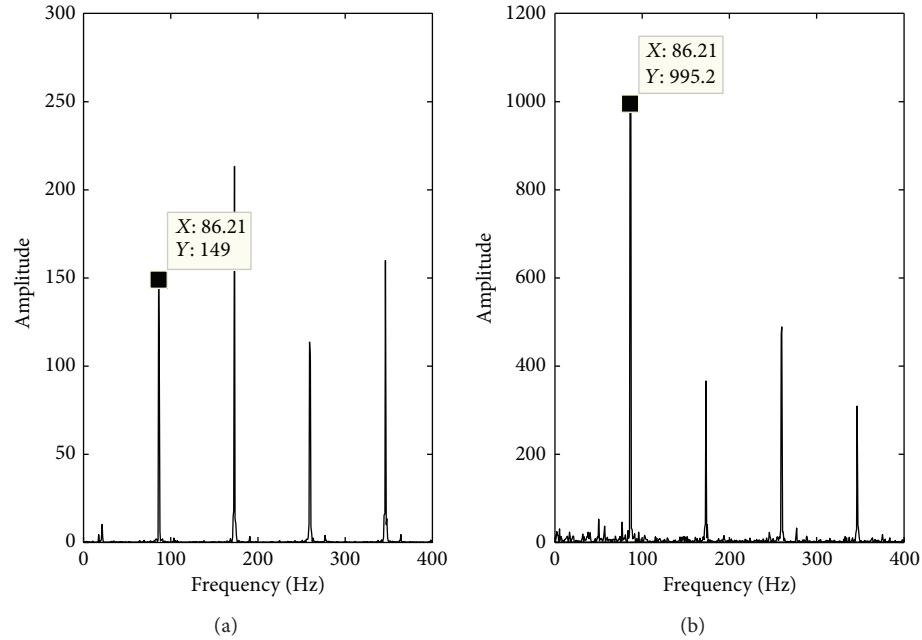


Figure 14: Envelope spectra of the vibration signal induced by outer race fault: (a) wavelet analysis method and (b) the proposed method.

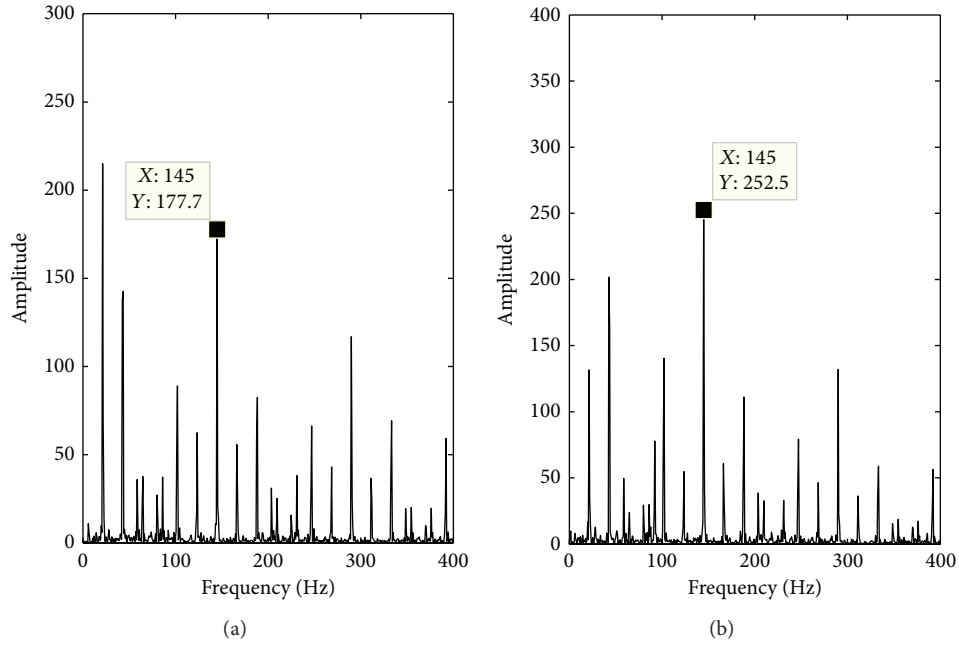


Figure 15: Envelope spectra of the vibration signal induced by inner race fault: (a) wavelet analysis method and (b) the proposed method.

If defect exists in the bearing, the magnitudes of faults frequencies will increase in the envelope spectrum. Therefore, they are often used to describe the fault severity and also employed to judge whether the fault features are enhanced. In this work, magnitudes of faults frequencies are chosen as the major indictors to reflect the fault features. As recorded and listed in Table 2, with the proposed method, magnitudes of faults frequencies are all enhanced and they become more easily distinguished from interfered noise for the experimental faults.

5. Conclusions

In order to improve the current wavelet-based method for fault detection of rolling bearings, the present paper developed a new way to improve vibration signals' smoothness

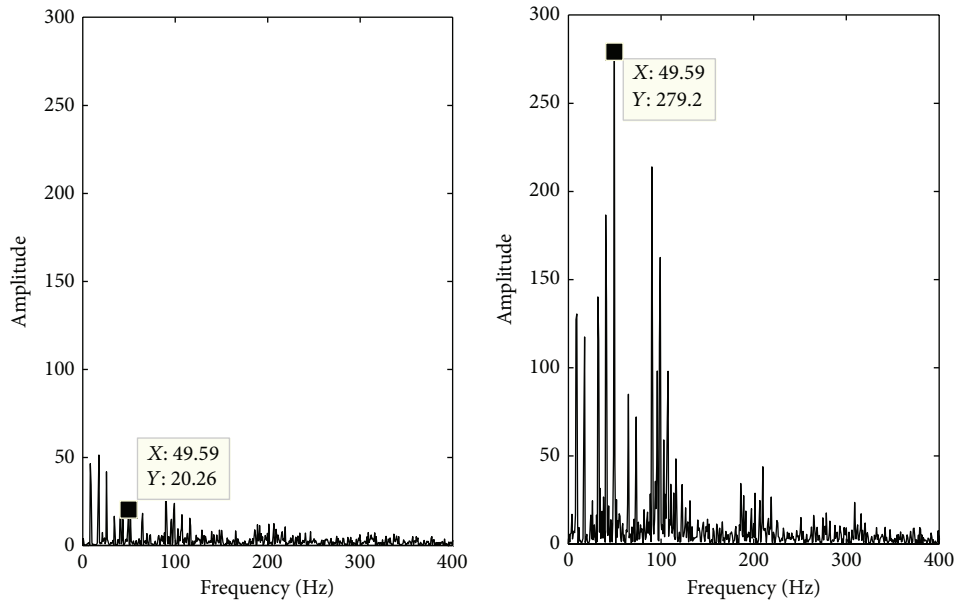


Figure 16: Envelope spectra of the vibration signal induced by roller element fault: (a) wavelet analysis method and (b) the proposed method.

Table 2: Major indictors for bearing fault features.

Defect frequency magnitude of	Traditional wavelet-based method	The proposed method
outer race	149	995.2
inner race	177.7	252.5
roller element	20.26	279.2

as well as their sparseness in wavelet domain by employing peak-based piecewise recombination and envelope demodulation. Through the proposed method, most noise mixed in vibration signals can be eliminated and the weak fault signals of rolling bearings can be enhanced and become more easily detectable. Therationality and validity of the methods are also proved by various experiments. It shows that, compared with traditional wavelet-based method, the proposed method has many advantages in bearing fault signals enhancement and detection and is very simple and easy to implement.

There is still room to improve the present work. The sensitivity of the method to noise and methods to make energy more concentrated are what will be studied in future work.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This project is partially supported by the National Natural Science Foundation of China (Grant no. 51375037) and the National Program on Key Basic Research Project (Grant no. 2012CB026000). The third author would also like to thank the support of China Fundamental Research Funds for the Central Universities (ZY1410) and the Public Hatching Platform for Recruited Talents of Beijing University of Chemical Technology.

References

[1] W. Zhou, T. G. Habetler, and R. G. Harley, "Bearing condition monitoring methods for electric machines: a general review," in *Proceedings of the IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics and Drives (SDEMPED '07)*, pp. 3–6, September 2007.

[2] H. Q. Wang and P. Chen, "Intelligent diagnosis method for a centrifugal pump using features of vibration signals," *Neural Computing and Applications*, vol. 18, no. 4, pp. 397–405, 2009.

[3] H. Wang and P. Chen, "Sequential diagnosis for rolling bearing using fuzzy neural network," in *Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM '08)*, pp. 56–61, Xi'an, China, August 2008.

[4] T. Mitoma, H. Wang, and P. Chen, "Fault diagnosis and condition surveillance for plant rotating machinery using partially-linearized neural network," *Computers and Industrial Engineering*, vol. 55, no. 4, pp. 783–794, 2008.

[5] P. D. McFadden and J. D. Smith, "Vibration monitoring of rolling element bearings by the high-frequency resonance technique—a review," *Tribology International*, vol. 17, no. 1, pp. 3–10, 1984.

[6] N. Tandon and A. Choudhury, "Review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings," *Tribology International*, vol. 32, no. 8, pp. 469–480, 1999.

[7] R. B. W. Heng and M. J. M. Nor, "Statistical analysis of sound and vibration signals for monitoring rolling element bearing condition," *Applied Acoustics*, vol. 53, no. 1–3, pp. 211–226, 1998.

[8] C. Junsheng, Y. Dejie, and Y. Yu, "The application of energy operator demodulation approach based on EMD in machinery fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 21, no. 2, pp. 668–677, 2007.

[9] R. B. Randall, J. Antoni, and S. Chobsaard, "The relationship between spectral correlation and envelope analysis in the diagnostics of bearing faults and other cyclostationary machine signals," *Mechanical Systems and Signal Processing*, vol. 15, no. 5, pp. 945–962, 2001.

[10] A. B. Ming, Z. Y. Qin, W. Zhang et al., "Spectrum auto-correlation analysis and its application to fault diagnosis of rolling element bearings," *Mechanical Systems and Signal Processing*, vol. 41, no. 1, pp. 141–154, 2013.

[11] D. Yu, J. Cheng, and Y. Yang, "Application of EMD method and Hilbert spectrum to the fault diagnosis of roller bearings," *Mechanical Systems and Signal Processing*, vol. 19, no. 2, pp. 259–270, 2005.

[12] V. K. Rai and A. R. Mohanty, "Bearing fault diagnosis using FFT of intrinsic mode functions in Hilbert-Huang transform," *Mechanical Systems and Signal Processing*, vol. 21, no. 6, pp. 2607–2615, 2007.

[13] Y. Lei, N. Li, J. Lin, and S. Wang, "Fault diagnosis of rotating machinery based on an adaptive ensemble empirical mode decomposition," *Sensors*, vol. 13, no. 12, pp. 16950–16964, 2013.

[14] J. Cheng, Y. Yang, and Y. Yang, "A rotating machinery fault diagnosis method based on local mean decomposition," *Digital Signal Processing*, vol. 22, no. 2, pp. 356–366, 2012.

[15] B. Chen, Z. He, X. Chen, H. Cao, G. Cai, and Y. Zi, "A demodulating approach based on local mean decomposition and its applications in mechanical fault diagnosis," *Measurement Science and Technology*, vol. 22, no. 5, Article ID 055704, 2011.

[16] S. Prabhakar, A. R. Mohanty, and A. S. Sekhar, "Application of discrete wavelet transform for detection of ball bearing race faults," *Tribology International*, vol. 35, no. 12, pp. 793–800, 2002.

[17] R. Yan, R. X. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: a review with applications," *Signal Processing*, vol. 96, part A, pp. 1–15, 2014.

[18] H. Wang and P. Chen, "A feature extraction method based on information theory for fault diagnosis of reciprocating machinery," *Sensors*, vol. 9, no. 4, pp. 2415–2436, 2009.

[19] K. Li, X. L. Ping, H. Q. Wang, and P. Chen, "Sequential fuzzy diagnosis method for motor roller bearing in variable operating conditions based on vibration analysis," *Sensors*, vol. 13, no. 6, pp. 8013–8041, 2013.

[20] R. Rubini and U. Meneghetti, "Application of the envelope and wavelet transform analyses for the diagnosis of incipient faults in ball bearings," *Mechanical Systems and Signal Processing*, vol. 15, no. 2, pp. 287–302, 2001.

[21] L. Gao, Z. Yang, L. Cai, H. Wang, and P. Chen, "Roller bearing fault diagnosis based on nonlinear redundant lifting wavelet packet analysis," *Sensors*, vol. 11, no. 1, pp. 260–277, 2011.

[22] Z. Yang, L. Cai, L. Gao, and H. Wang, "Adaptive redundant lifting wavelet transform based on fitting for fault feature extraction of roller bearings," *Sensors*, vol. 12, no. 4, pp. 4381–4398, 2012.

[23] Z. He, "Peak transform for efficient image representation and coding," *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1741–1754, 2007.

[24] S. Anila and N. Devarajan, "The usage of peak transform for image compression," *International Journal of Computer Systems Science & Engineering*, vol. 2, no. 11, pp. 6308–6316, 2010.

[25] D. Ho and R. B. Randall, "Optimization of bearing diagnostic techniques using simulated and actual bearing fault signals," *Mechanical Systems and Signal Processing*, vol. 14, no. 5, pp. 763–788, 2000.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 781583, 10 pages
http://dx.doi.org/10.1155/2014/781583

Research Article

Bearing Condition Recognition and Degradation Assessment under Varying Running Conditions Using NPE and SOM

Shaohui Zhang[1] and Weihua Li[1,2]

[1] *School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510640, China*
[2] *State Key Laboratory for Manufacturing System Engineering, Xi'an Jiaotong University, Xi'an 710049, China*

Correspondence should be addressed to Weihua Li; whlee@scut.edu.cn

Received 7 March 2014; Accepted 3 April 2014; Published 5 May 2014

Academic Editor: Ruqiang Yan

Manifold learning methods have been widely used in machine condition monitoring and fault diagnosis. However, the results reported in these studies focus on the machine faults under stable loading and rotational speeds, which cannot interpret the practical machine running. Rotating machine is always running under variable speeds and loading, which makes the vibration signal more complicated. To address such concern, the NPE (neighborhood preserving embedding) is applied for bearing fault classification. Compared with other algorithms (PCA, LPP, LDA, and ISOP), the NPE performs well in feature extraction. Since the traditional time domain signal denoising is time consuming and memory consuming, we denoise the signal features directly in feature space. Furthermore, NPE and SOM (self-organizing map) are combined to assess the bearing degradation performance. Simulation and experiment results validate the effectiveness of the proposed method.

1. Introduction

Bearings are among the most essential components in rotating machinery. They are frequently operated under high loading and severe conditions. And the defects often occur gradually on the bearing. In order to prevent unexpected bearing failures, many fault diagnosis methods have been explored for faults detection, such as temperature monitoring, oil analysis, and vibration analysis. Among them, vibration signal-based method has been extensively used for bearing condition monitoring. The traditional vibration signal processing methods include time domain analysis [1], frequency domain analysis (such as FFT transform and envelope demodulation [2]), and time-frequency domain analysis (such as wavelet transform [3] and WV distribution [4]). Many features can be extracted from vibration data using these methods. Feature extraction is implemented to serve as a preprocessor for the fault diagnosis and the performance assessment. The classical feature extraction methods, such as PCA (principal component analysis) [5] and LDA (linear discriminant analysis) [6], have been successfully applied in fault diagnosis.

However, PCA and LDA are preferable in applications where the data space is linear. As opposed to these techniques, some dimensionality reduction techniques were proposed for processing nonlinear data structure in recent years, such as NPE (neighbor preserving embedding) [7], LPP (locality preserving projections) [8], and ISOP (isometric projection) [9]. And these methods have been effectively used in machine condition monitoring and fault diagnosis. Yu [10] proposed a local and nonlocal preserving projection (LNPP) algorithm for machine fault diagnosis and performance prognostics. Yang et al. [11] adopted the principal manifold learning to reduce the noise of nonlinear time series. Jiang et al. [12] developed supervised Laplacian eigenmaps technique for gearbox fault classification. Li and Zhang [13] used supervised locally linear embedding projection for bearing fault diagnosis.

One of the primary difficulties in bearing condition monitoring is how to eliminate the noise influence. Generally, vibration signals are sampled as longtime series in order to improve the frequency resolution. The traditional denoising methods, like SVD (singular value decomposition) [14] and wavelet [15], are time consuming and memory consuming.

When various features are extracted from vibration data, the noise contained in the data is also transferred to the features. Therefore, we think that denoising these features directly can speed up the computation and also save the memory space.

In addition, studies of bearing fault diagnosis focused on the bearing faults under stable loading and rotational speeds, which cannot interpret the machine running conditions. The bearings were always running under variable loading and speeds, and the actual vibration signal is more complicated. How to process these complicated vibration signals is another key issue to perform bearing health assessment.

The objective of this work is to explore the effectiveness of unsupervised NPE algorithm on bearing fault diagnosis and degradation assessment. The experimental results illustrate that the proposed scheme is capable of identifying different fault modes and evaluating degradation performance.

The rest of the paper is organized as follows. Section 2 presents the basic theory of NPE algorithm briefly. In Section 3, the feature-denoising algorithm is presented and the NPE approach is used to recognize the bearing vibration signals measured under variable loading and speeds. In Section 4, NPE is combined with SOM (self-organizing map) to describe the bearing defect propagation. The conclusion in Section 5 closes the paper.

2. Neighborhood Preserving Embedding (NPE)

Locally linear embedding (LLE) does not require an iterative algorithm and just a few parameters need to be set, which leads to its application in fault diagnosis. However, the performance of the LLE is sensitive to the selection of the nearest neighbors. The NPE can avoid this disadvantage, since it is less sensitive to outliers than LLE. NPE aims at preserving the local manifold structure after dimension reduction, and it is a linear approximation of the LLE.

Given a set of N samples assembled in a matrix $\mathbf{X} = [x_1, x_2, x_3, \ldots, x_N]$, size $M \times N$, a transformation matrix $\mathbf{A}$ can be constructed to project these N samples to images assembled in a matrix $\mathbf{Y} = [y_1, y_2, y_3, \ldots, y_N]$, size $d \times N$ ($d \ll M$), where the rth column vector of $\mathbf{Y}$ corresponds to that of $\mathbf{X}$, respectively.

(1) Constructing an adjacency graph: calculate Euclidean distance between samples x_i and x_j, and use k-nearest neighbor to construct the adjacency graph $\mathbf{G}$. The distance $d(x_i, x_j)$ represents the edge connecting x_i and x_j, as

$$d\left(x_i, x_j\right) = \left\|x_i - x_j\right\|. \tag{1}$$

(2) Computing the weights: for each sample, it can be represented as a linear combination of the neighbors, and the weight matrix reflects the coefficients. In this step, we can compute the corresponding weight w_{ij} by minimizing the following weighted cost function $\varepsilon(\mathbf{w})$:

$$\varepsilon\left(\mathbf{w}\right) = \min \left\| x_i - \sum_{j=1}^{k} w_{ij} x_i^j \right\|^2, \tag{2}$$

where ε is the reconstruction error, w_{ij} is the weight of the jth neighbor of data x_i, with the constraint $\sum_{j=1}^{k} w_{ij} = 1$, and x_i^j is the jth neighbor of data x_i. Let $(w_{i1}, w_{i2}, \ldots, w_{ik})$ be a weight vector and obtain N-D vector $\mathbf{W}_i$ by adding zero for the nonneighbors. All these vectors are used to construct a matrix $\mathbf{W}$, size $N \times N$. Equation (2) can be rewritten as

$$\varepsilon\left(\mathbf{W}_i\right) = \min \left\| x_i - \mathbf{X}\mathbf{W}_i^T \right\|^2. \tag{3}$$

Therefore,

$$\begin{aligned} \varepsilon\left(\mathbf{W}\right) &= \min \left\{ \sum_{i=1}^{N} \left\| x_i - \mathbf{X}\mathbf{W}_i^T \right\|^2 \right\} \\ &= \min \left\{ \mathrm{tr}\left(\mathbf{X}\mathbf{M}\mathbf{X}^T\right) \right\}, \end{aligned} \tag{4}$$

where tr is the trace of $\mathbf{X}\mathbf{M}\mathbf{X}^T$, $\mathbf{M} = (\mathbf{I} - \mathbf{W}^T)^T(\mathbf{I} - \mathbf{W}^T)$, and $\mathbf{I}$ is an $N \times N$ identity matrix.

Suppose the transformation is $\mathbf{Y} = \mathbf{A}^T\mathbf{X}$. For the purpose of removing an arbitrary scaling factor in the projection, we impose a constraint function as follows:

$$\mathbf{Y}\mathbf{Y}^T = 1 \Longrightarrow \mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A} = 1. \tag{5}$$

(3) Computing the projections and combining (4) and (5) yield the following generalized eigenvector problem:

$$\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A} = \lambda\mathbf{X}\mathbf{X}^T\mathbf{A}. \tag{6}$$

$a_0, \ldots, a_{d-1}$ are arranged according to their corresponding eigenvalues $\lambda_0 \le \lambda_1 \cdots \le \lambda_{d-1}$. Therefore, the embedding is performed as follows:

$$x_i \longrightarrow y_i = \mathbf{A}^T x_i, \tag{7}$$

where y_i is a d dimensional vector and $\mathbf{A} = (a_0, a_1, \ldots, a_{d-1})$, size $M \times d$.

It should be pointed out that "labeled" data is often used for supervised learning while "unlabeled" data for unsupervised learning, and NPE can adopt both of them for learning process. The class information can be utilized to get a better dimensionality reduction. But in practical applications, it is often difficult to collect labeled instances. However, unlabeled data may be relatively easy to acquire. The focus of this paper is to demonstrate the effectiveness of NPE in bearing fault classification and degradation assessment, without class information.

3. Simulation and Experiments

3.1. Setup. In this section, the NPE is used to process the simulation signals and the bearing vibration data. Then, the 1NN (1-nearest neighbor) method is adopted to classify different bearing fault, because it is the simplest classification algorithm which can be used even with few samples, and it works very well in low dimensions for complex decision surfaces. The procedure of machine condition recognition and health assessment is illustrated in Figure 1.

In order to highlight the effectiveness of the NPE algorithm, the results of NPE are compared with those of unsupervised learning (LPP and PCA) and supervised learning (LDA, ISOP, and supervised NPE as SNPE), respectively.

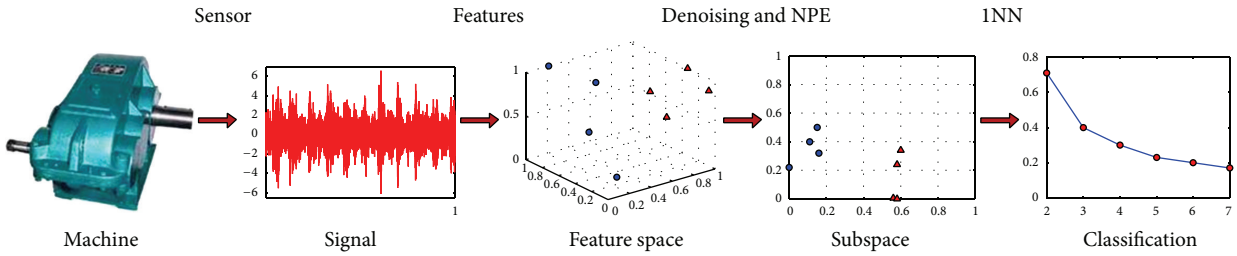


Figure 1: The process of machine condition recognition and health assessment.

Case 1 (*simulation*). To describe the waveform generated by a rolling element bearing under the constant radial load with a single localized defect, the vibration signature can be expressed as (8) in [16]:

$$y(t) = y_d(t)\, y_q(t)\, y_r(t)\, y_e(t) + y_n(t), \tag{8}$$

where y_d is a series of impulses at bearing fault frequency, y_q is the bearing radial load distribution, y_r represents the bearing induced resonant frequency, y_e reflects the exponential decay due to damping, and y_n is the noise.

The rotational speed was set at 1100 rpm, and the vibration signals were measured at a constant sampling rate 120 kHz; the duration of each vibration signal was 42 seconds. The fault frequency was placed at 133 Hz and impact amplitudes were set as 0, 1, 1.5, and 2, which represented the normal, slight, moderate, and severe fault, respectively. White Gaussian noise with SNR = −2 dB was added to the signal.

Case 2 (*bearing conditions recognition under varying working conditions*). The data used in this work were obtained from the Case Western Reserve University Bearing Data Centre [17]. All the experiments were repeated for different loading conditions: 1, 2, and 3 hp at rotational speeds ranging from 1730 rpm to 1772 rpm. The data were sampled at a rate of 48 kHz and the duration of each vibration signal was 10 seconds. Defects were introduced into the rolling element bearing at drive-end by using electrodischarge machining in the following configurations:

(i) healthy bearing being used as a baseline,

(ii) inner race defect (0.007 inch, 0.014 inch, 0.021 inch in diameter, and 0.011 inch in depth),

(iii) outer race defect (0.007 inch, 0.014 inch, 0.021 inch in diameter, and 0.011 inch in depth),

(iv) ball defect (0.007 inch, 0.014 inch, 0.021 inch in diameter, and 0.011 inch in depth).

So there are totally 10 kinds of bearing conditions, and each defect has three different levels under varying loading and speeds, and there are 600 vibration data sampled in total (60 datasets per condition). Then 20 features (including time domain and frequency domain features, as in Table 1) were extracted from the data. Therefore, the original feature data is of size (600 × 20) in high-dimension space.

Table 1: Bearing feature generation.

Time domain features	Frequency domain features
Mean square	A_{f_n} amplitude at rotating frequency f_n
Kurtosis	A_{2f_n} amplitude at $2f_n$
Mean	A_{BPFI} amplitude at BPFI
Skewness	A_{2BPFI} amplitude at 2BPFI
Peak value	A_{BPFO} amplitude at BPFO
Rms	A_{2BPFO} amplitude at 2BPFO
Waveform factor	A_{FTF} amplitude at FTF
Crest factor	A_{2FTF} amplitude at 2FTF
Impulse factor	A_{BSF} amplitude at BSF
Clearance factor	A_{2BSF} amplitude at 2BSF

Case 3 (*bearing degradation assessment using NPE and SOM*). The data under normal state and ball defects (0.007 inch, 0.014 inch, 0.021 inch in diameter, and 0.011 inch in depth) were selected to analyze the bearing degradation trend.

3.2. Simulation Results Analysis. The simulation signal under different impacts is shown in Figure 2. It can be seen that there is little difference between the normal signal and slight fault one, even the moderate fault, due to the impact of noise. As for the severe fault bearing signal, the impact amplitude is larger than those of others. Each vibration signal was divided into forty segments and the corresponding features as listed in Table 1 were extracted, and there were 160 20-D feature data sets obtained in total. The feature curves before and after denoising are shown in Figure 3. It can be observed that denoising features in the feature space can reduce the margin fluctuation of the curves, which is beneficial to the fault classification, especially for features as mean, skewness, impulse factor, clearance factor, and most of frequency features.

For verifying the generalization performance of the proposed model, 2 random subsets with 25% and 50% of the simulation data were used for training, and the rest for testing. This was aimed at examining the performance of the NPE dealing with data sets for which it was not sufficiently trained. The training dataset was used to obtain the transformation matrix **A**. Then the testing data were input into the learning machine to construct the feature subspace and used the 1NN to calculate misclassification rate. This process was repeated for 20 times, and the result was the

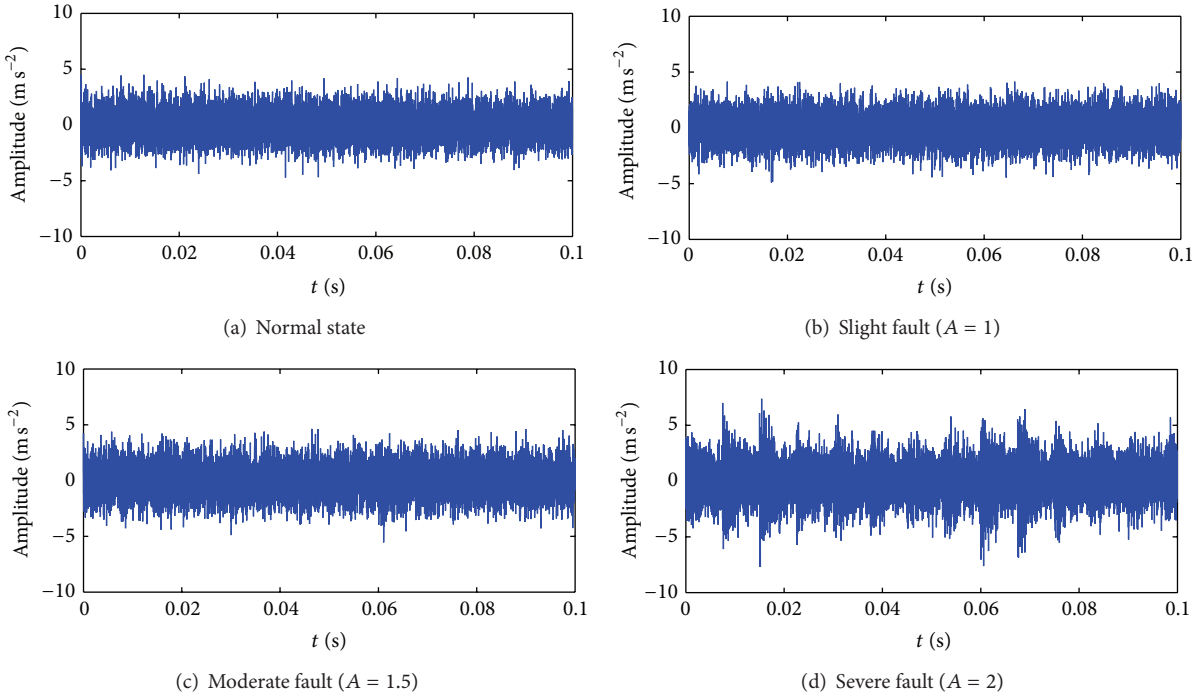


Figure 2: Simulation of vibration signals under different impacts.

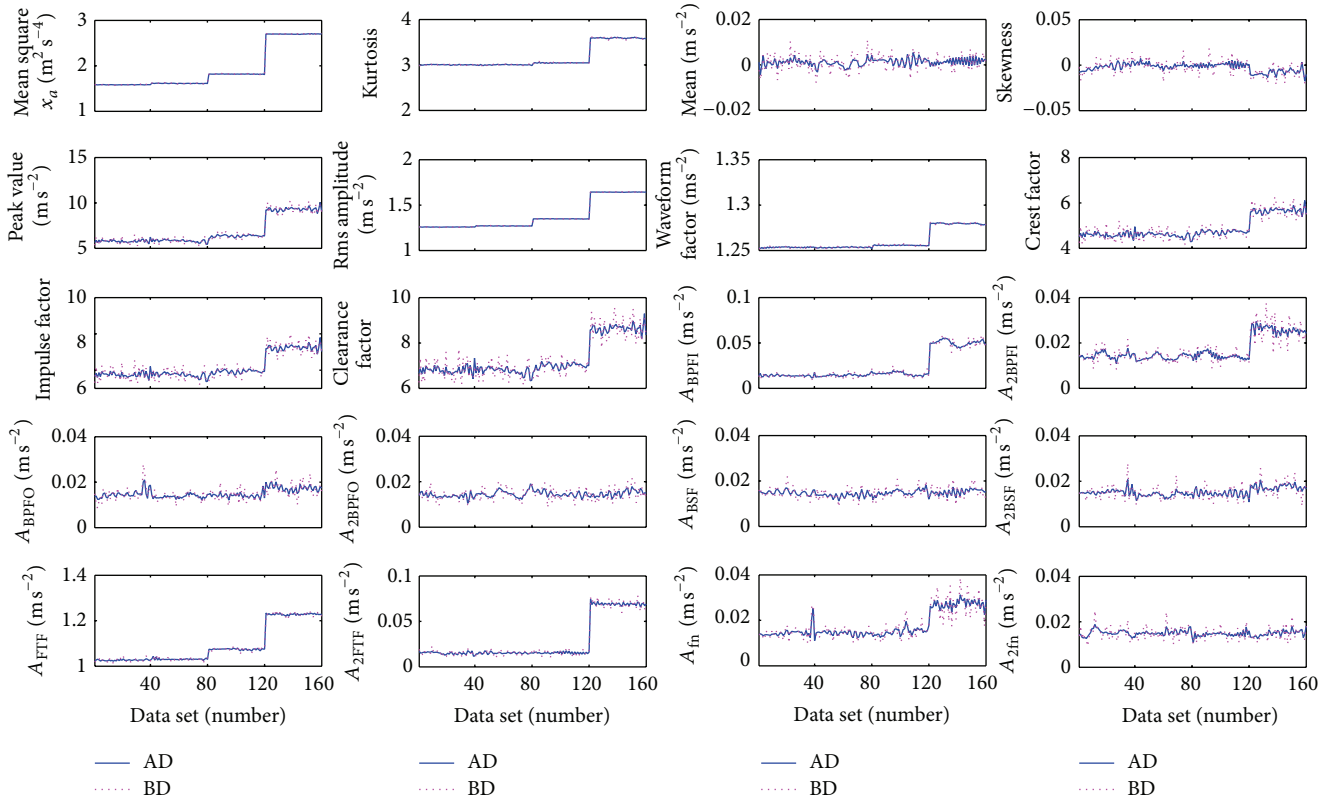
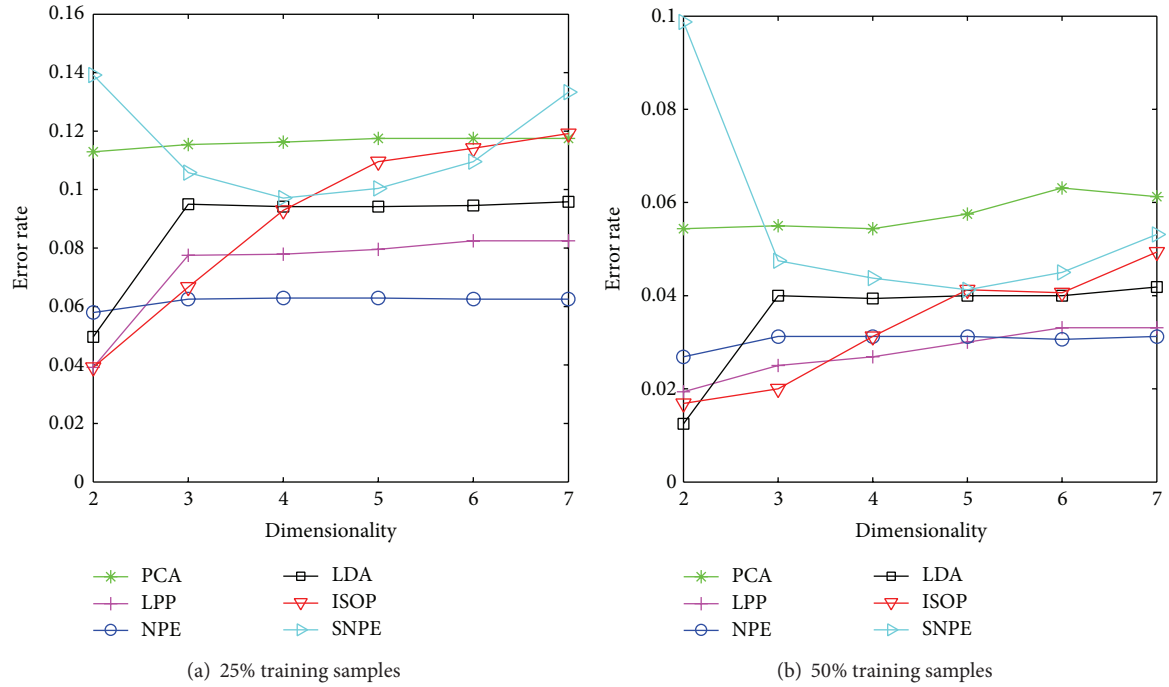
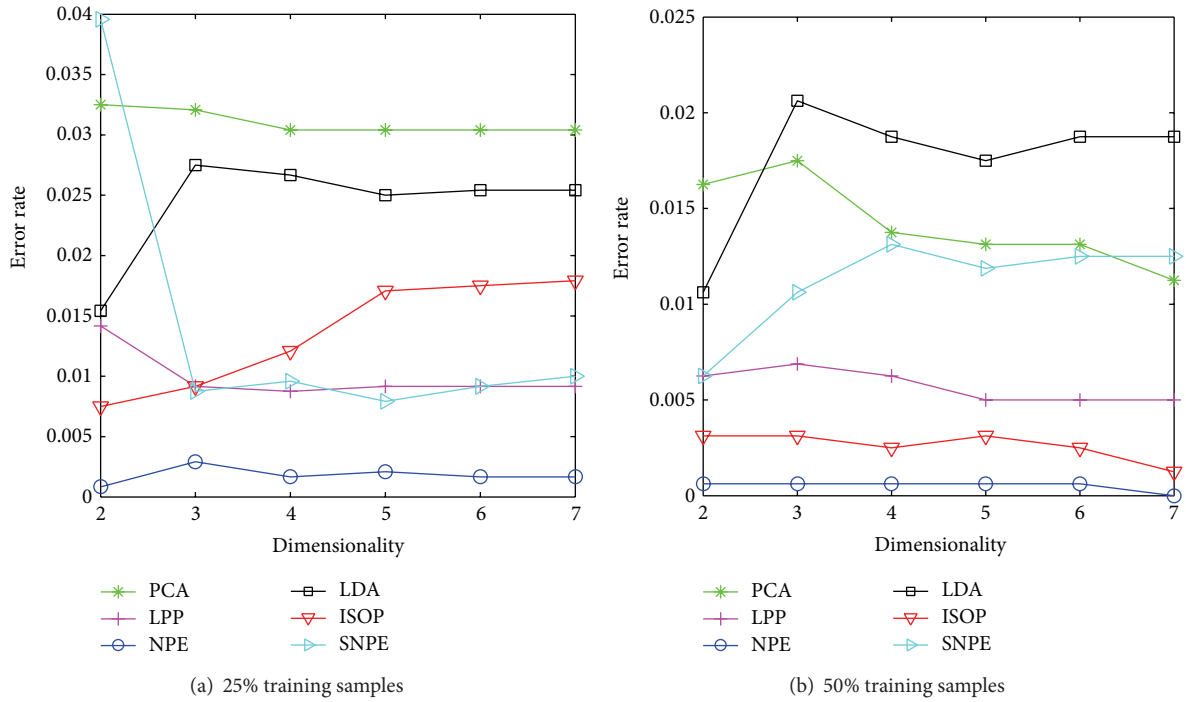


Figure 3: Features variation before and after denoising (BD: before denoising and AD: after denoising).



(a) 25% training samples (b) 50% training samples

FIGURE 4: Simulation result: the error rate of different algorithms without denoising.



(a) 25% training samples (b) 50% training samples

FIGURE 5: Simulation result: the error rate of different algorithms with denoising.

average of 20 processes. The error rate versus dimension is shown in Figure 4.

As can be seen from Figure 4(a), when 25% samples were used for training, the misclassification of unsupervised NPE is lower than that of others. In 2-dimension space, the LPP, LDA, and ISOP perform better, and their error rates are 3.912%, 4.958%, and 3.917%, respectively. In Figure 4(b), given 50% samples for training, both of algorithms can achieve good results, and the error rates of them are less than 10%. The aforementioned learning methods perform better in 2D space, and the misclassification rate of SNPE is the maximum, 9.875%. However, the performance of NPE

is stable and acceptable; the error rate is ranging from 5.792% to 6.292% (25% training samples) or 2.688%–3.125% (50% training samples). When the dimension increases, it is superior to other algorithms.

To speed up computation and save memory space, the feature samples are denoised directly using singular value decomposition (SVD). The steps of SVD for noise reduction are as follows.

(1) The vibration signal $\mathbf{X}$ can be transformed into an $n \times m$ matrix $\mathbf{Q}$ by using phase space reconstruction with the time delay of "one unit." The SVD of $\mathbf{Q}$ is a factorization of the formula $\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ is an $n \times n$ real or complex unitary matrix, $\mathbf{V}^T$ (the conjugate transpose of $\mathbf{V}$) is an $m \times m$ real or complex unitary matrix, and $\mathbf{\Sigma}$ is a diagonal matrix of size $n \times m$, and its diagonal values are the singular values of matrix $\mathbf{Q}$. The singular value reflects the energy concentration between signals and noises. The greater the value is, the less influence the noise affects. Keep great values and set all the small values as zero, which will effectively improve signal-to-noise ratio and obtain a diagonal matrix $\mathbf{\Sigma}'$.

(2) Calculate space matrix again as $\mathbf{Q}' = \mathbf{U}\mathbf{\Sigma}'\mathbf{V}^T$ and restore signal $\mathbf{X}'$ to one-dimensional series after noise reduction.

After feature denoising, the learning and classification were repeated using NPE and other algorithms. The error rates of different algorithms with denoising feature samples are shown in Figure 5.

Comparing Figure 4 with Figure 5, it can be observed that denoising feature sets can depress the misclassification effectively. Both of the algorithms can achieve good results, and all the error rates are less than 4%. As can be seen from Figure 5, the performance of NPE is stable and the misclassification of NPE is the lowest of these algorithms; its error rate is ranging from 0.0833% to 0.2917% (25% training samples) and 0%–0.0625% (50% training samples). The largest error rate is of SNPE, 3.958% with 25% training samples, while that is 2.063% of LDA with 50% training samples.

3.3. Experiment Results and Discussion. Bearing datasets in Case 2 were used to validate the proposed approach in bearing condition recognition.

For verifying the generalization performance of the proposed model, 2 random subsets with 30% and 50% of the data were used for training, and the rest for testing. This process was repeated for 20 times, and the result was the average of 20 processes. The error rate versus dimension is revealed in Figure 6.

As depicted in Figure 6, the misclassification rate decreased with the dimension increasing and the training samples growing, and the supervised methods are superior to the unsupervised ones. It is obvious that the misclassification of LDA, ISOP, and SNPE is lower than those of unsupervised methods. But in reality, it is often difficult or expensive to collect labeled instances. However, unlabeled data may be relatively easy to acquire. The unsupervised learning methods are more appropriate for practical application.

The NPE achieves the 90% classification correctness (error rate 9.767%, 50% training samples) in the 4-dimension feature space. It can be seen that, for multifault classification, the dimension of feature space affects the classifying results. In 2D and 3D space, most of the methods cannot get the good classification, except for the SNPE (error rate 9.917%, 50% training samples in 3D space).

In order to validate the feature denoising, we denoised features directly by SVD, and the results are shown in Figure 7.

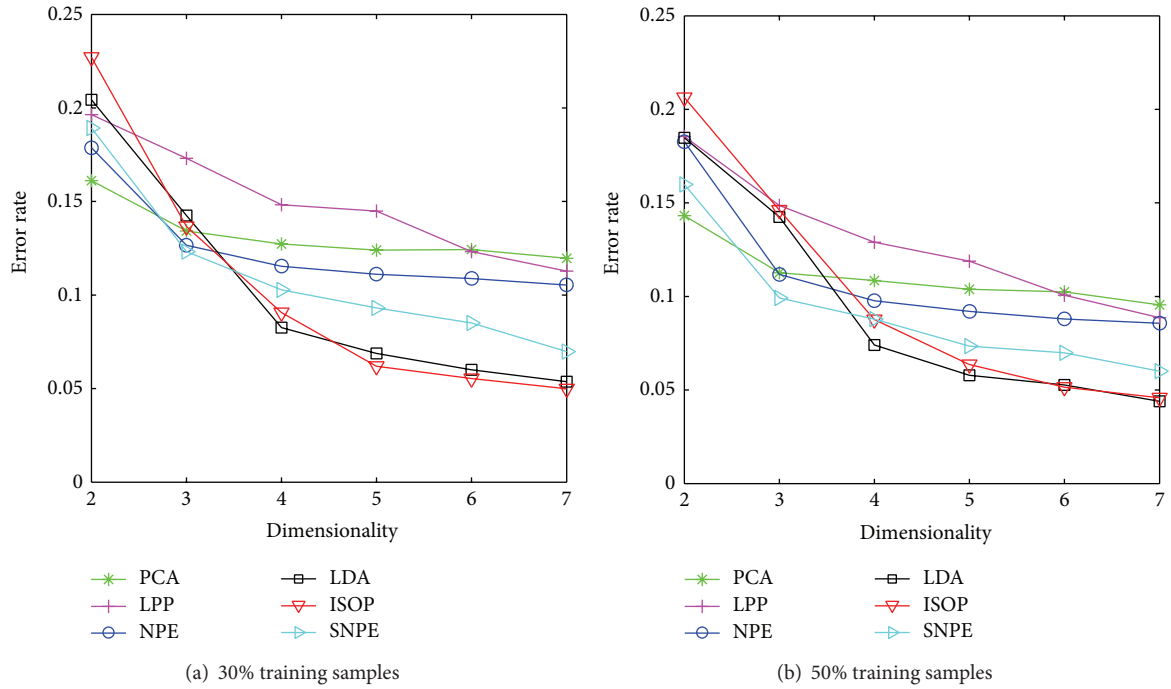
Comparing Figure 6 with Figure 7, it can be observed that denoising feature sets can decrease the misclassification rate effectively. Both of the algorithms can achieve better results than those before denoising, and all the error rates are less than 10% when the dimension is more than 2. As can be seen from Figure 7, the performance of NPE is also agreeable and acceptable. The misclassification of NPE is the lowest among unsupervised algorithms (PCA, LPP, and NPE) when the subspace dimension exceeds 3, and its error rate is ranging from 4.452% to 6.381% (30% training samples) and 3.45% to 5.88% (50% training samples). Whereas in 3D feature space, it achieves the best classification result (the error rate 6.381%) with only 30% training samples involved in learning.

4. Bearing Degradation Performance Assessment Using NPE and SOM

In this section, the bearing defect propagation was investigated by using SOM. Rolling element bearings with ball defect (normal, fault size with 0.007 inch, 0.014 inch and 0.021 inch in diameter, and 0.011 inch in depth) under different loadings (1 hp, 2 hp, and 3 hp) were used to implement the degradation assessment, 240 datasets in total. Firstly, the unsupervised NPE was used for feature extraction. Then the MQE (minimum quantization error) of the SOM can be used to observe the bearing degradation trend.

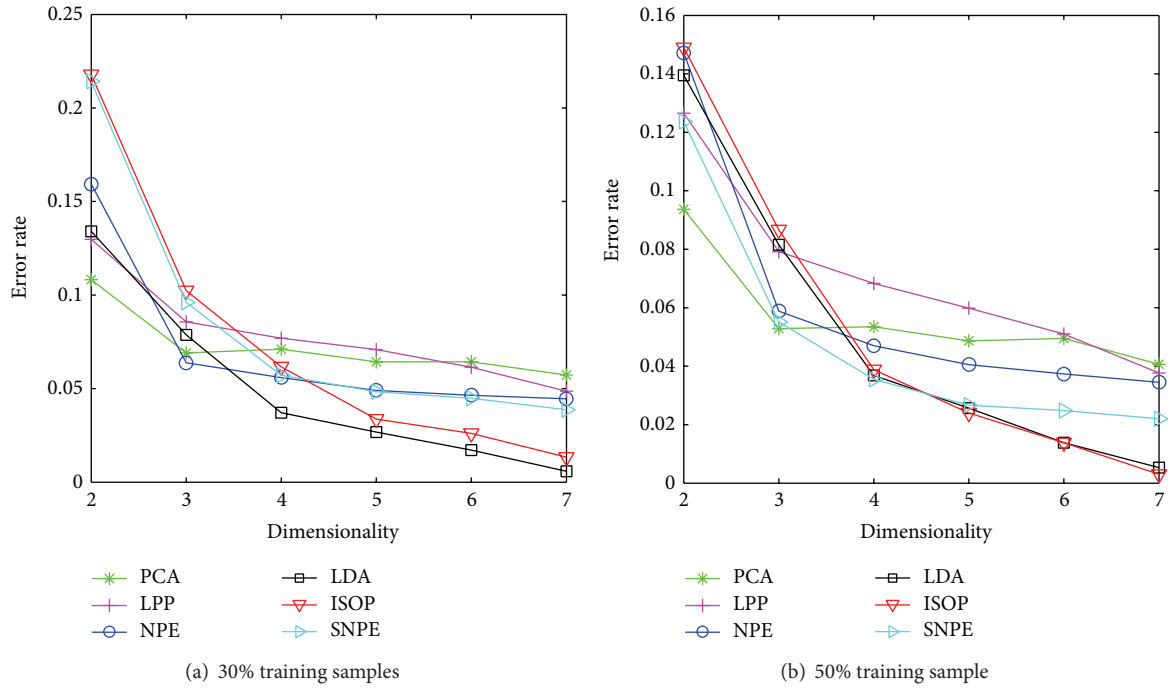
SOM (self-organizing map) was developed by Kohonen [18], and it is a general unsupervised tool for clustering. It consists of neurons located on a regular, usually 2D grid of map units. And for similar samples in the input space, they can be mapped to the neighbor neurons in the output space. SOM has been proven useful in gearbox condition monitoring [19], degradation assessment of rolling element bearing [20], and so on.

In fact, it is often difficult to collect datasets which can represent the whole failure space. Meanwhile, healthy samples can be obtained relatively easy, which can be used to characterize the normal state. And the deviation from the normal feature space can reflect degradation detection. A quantitative degradation index as the MQE of SOM can be obtained by depending on distance between the normal state and the current process, which can be normalized by converting the MQE into confidence value (CV) ranging from 0 to 1. At first, only the healthy data are needed for training SOM model. The new sample is input into the model and compared with the weight vectors of all map units,



(a) 30% training samples (b) 50% training samples

Figure 6: Bearing fault classification: the error rate of different algorithms without denoising.



(a) 30% training samples (b) 50% training sample

Figure 7: Bearing fault classification: the error rate of different algorithms with denoising.

and when the smallest difference exceeds a predetermined threshold, this sample is probably in a fault situation.

Quantization error shows how accurately the neurons of the trained network respond to the given input samples and is also the average distance between the data vector x_i and the BMU (best matching unit). The MQE could be calculated according to (9), and more detail information about the algorithm can be found in [20], as follows:

$$\mathrm{MQE}(i) = \left\| x_i - h_{\mathrm{BMU}} \right\|, \tag{9}$$

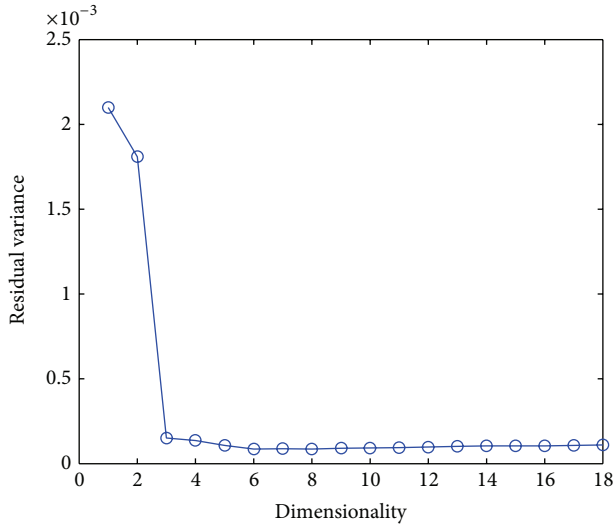


Figure 8: Residual curve of bearing running state.

where h_{BMU} represent the BMU of healthy data, and x_i denotes the test samples.

To investigate the bearing degradation performance, we need to represent the bearing running tendency in a low dimensional space. Therefore, NPE was used to decrease the dimension of feature space. The intrinsic dimension of the original feature space describes how many variables are needed to represent the bearing running. There are two kinds of methods to estimate the intrinsic dimension. One is the eigenvalue method, such as PCA, which determines the intrinsic dimension through the number of eigenvalues greater than a given threshold; however, it failed on the nonlinear manifold. And the other is the geometric one, such as ISOMAP (isometric feature mapping), which exploits the intrinsic geometry based on the nearest neighbor distances. It provides residual error curves that can be "eyeballed" to estimate the intrinsic dimension [21], and residual variance decreases with added dimensions. The intrinsic dimension of the data can be estimated by finding the "elbow" at which the residual curve ceases to decrease significantly with dimension increases, as shown in Figure 8. It can be seen that the "elbow" point is at the place where the dimension is three. It means that the 3D subspace can be used to describe bearing running trend.

Therefore, three new features can be obtained by the NPE from the original feature sets. To eliminate the noise influence, the original features were denoised, and the variation of these features was described in Figure 9(b), in comparison with that before denoising (as shown in Figure 9(a)). It can be seen that, after feature denoising, the intraclass samples gathered together and the interclass ones were separated relatively.

A random subset with 50% of the data in Case 3 was selected for training, while the remaining for testing. At first, the NPE was trained to obtain the transformation matrix **A**. Then the testing data were input into the learning machine to construct the feature subspace. To validate the MQE's capability of degradation detection, SOM was firstly trained by the selected normal datasets from subspace. Then the test data in subspace was input to the learnt SOM, and the confidence value (MQE) was calculated to measure the deviation from the normal state. This process was repeated for 20 times, and the result was the average of 20 processes.

We compared the MQE of the original feature space with those of NPE feature subspace. Figure 10 shows the MQE curves of the original features, the NPE features, and the NPE features with denoising, respectively.

Comparing Figure 10(a) with Figure 10(b), it can be observed that the bearing degradation tendency can be described in the NPE subspace, validating that the intrinsic dimension is determined correctly. However, it is still unable to distinguish the ball defect BD014 (0.014 inch) from defect BD021 (0.021 inch). The confidence value of BD021 (1 hp, 1772 rpm) data is lower than that of BD014 (1 hp, 2 hp, and 3 hp), due to the noise. Even the confidence values of the same degradation state BD014 were varying greatly under different working conditions.

Therefore, the original features were denoised in order to eliminate the noise effect on vibration features, where the result was shown in Figure 10(c). It can be seen that the confidence value varies with different state obviously, especially for the moderate defect BD014 and the severe defect BD021. The worse the rolling element degraded, the bigger the confidence value (MQE) increased. At the same time, the confidence value of the same state fluctuated narrowly, especially for BD014, while the mean value of BD021 (1 hp, 1772 rpm) was almost the same as that of BD014 (3 hp, 1730 rpm). The results indicated that noise reduction facilitates a reliable rolling bearing performance prediction, and the proposed NPE-MQE can effectively assess the bearing degradation performance.

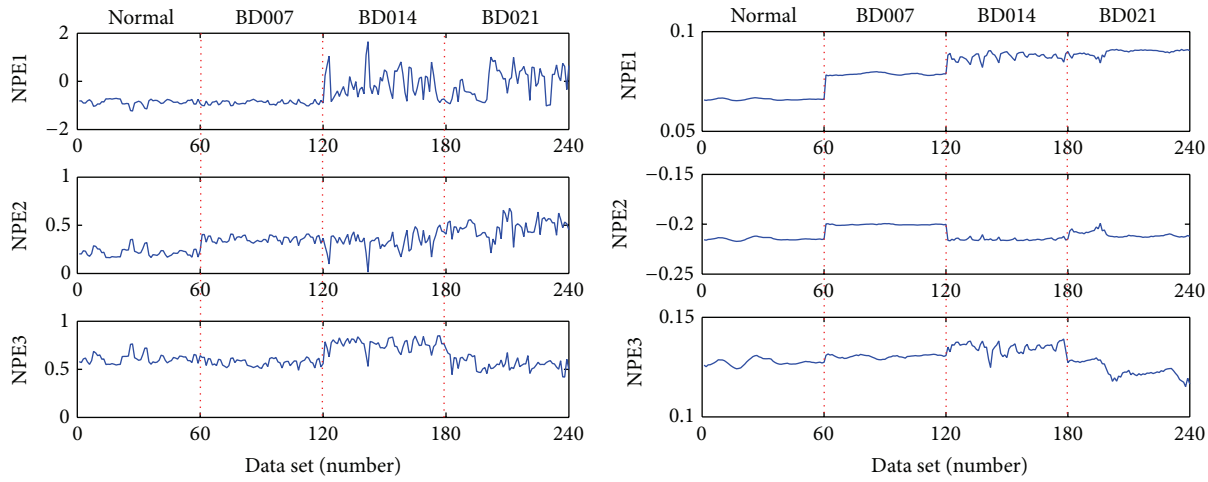
5. Conclusions

To investigate the fault recognition under varying working conditions, the NPE was adopted to perform dimension reduction and 1NN was used to classify different faults. Furthermore, the NPE and SOM were combined together for bearing degradation performance evaluation. The work in this study can be summarized as follows.

(1) Denoising features can speed up the computation, save memory space, and improve the recognition accuracy effectively, which is very helpful for fault classification.

(2) Simulation and experiment results demonstrate that the NPE is capable of extracting discriminative features, even lacking training samples. It is beneficial to both fault classification and degradation assessment.

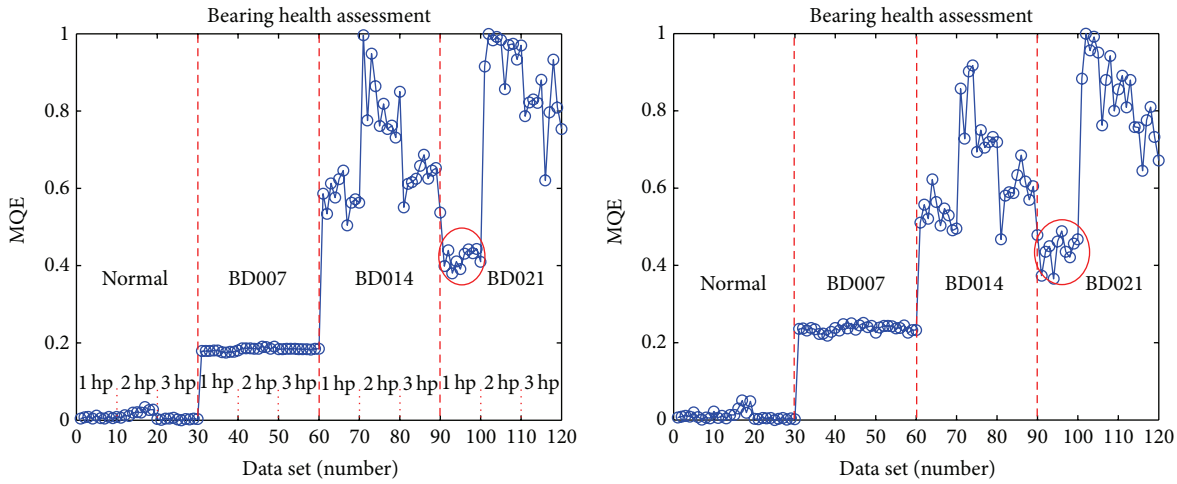
(3) The proposed NPE-MQE method can be used to assess the bearing degradation performance, and the confidence value can depict the bearing degradation process efficaciously.



(a) NPE features without denoising

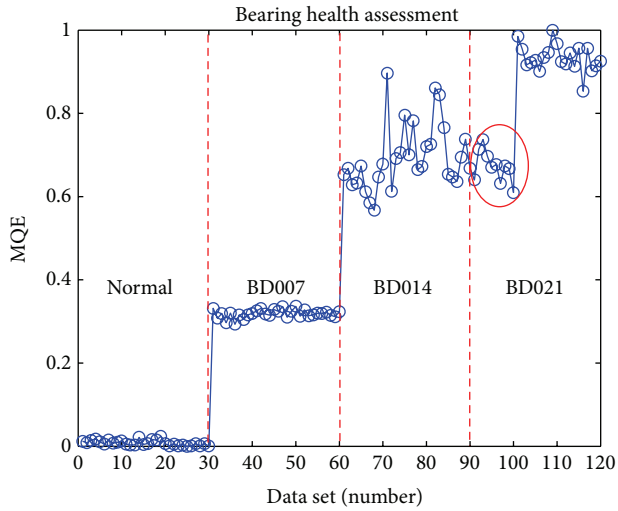
(b) NPE features with denoising

Figure 9: New features extracted by NPE.



(a) The MQE of original features

(b) The MQE of NPE features



(c) The MQE of NPE features with denoising

Figure 10: The MQE variation in time.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work described in this paper is supported in part by the National Natural Science Foundation of China (Grant no. 51075150) and State Key Laboratory for Manufacturing System Engineering (SKLMS2014008).

References

[1] F. Gao and Y. Lu, "A Kalman-filter based time-domain analysis for structural damage diagnosis with noisy signals," *Journal of Sound and Vibration*, vol. 297, no. 3–5, pp. 916–930, 2006.

[2] H. Zoubek, S. Villwock, and M. Pacas, "Frequency response analysis for rolling-bearing damage diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 55, no. 12, pp. 4270–4276, 2008.

[3] J. Kim, D. E. Welcome, R. G. Dong, W. J. Song, and C. Hayden, "Time-frequency characterization of hand-transmitted, impulsive vibrations using analytic wavelet transform," *Journal of Sound and Vibration*, vol. 308, no. 1-2, pp. 98–111, 2007.

[4] W. J. Staszewski, K. Worden, and G. R. Tomlinson, "Time-frequency analysis in gearbox fault detection using the Wigner-Ville distribution and pattern recognition," *Mechanical Systems and Signal Processing*, vol. 11, no. 5, pp. 673–692, 1997.

[5] I. Jolliffe, *Principal Component Analysis*, Springer, New York, NY, USA, 1986.

[6] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, NY, USA, 1973.

[7] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 1208–1213, October 2005.

[8] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*, vol. 16, pp. 153–160, MIT Press, Cambridge, Mass, USA, 2004.

[9] D. Cai, X. He, and J. Han, "Isometric projection," in *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI '07)*, pp. 528–533, July 2007.

[10] J. Yu, "Local and nonlocal preserving projection for bearing defect classification and performance assessment," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 5, pp. 2363–2376, 2012.

[11] J. Yang, J. Xu, D. Yang, and M. Li, "Noise reduction method for nonlinear time series based on principal manifold learning and its application to fault diagnosis," *Chinese Journal of Mechanical Engineering*, vol. 42, no. 8, pp. 154–158, 2006.

[12] Q. Jiang, M. Jia, J. Hu, and F. Xu, "Machinery fault diagnosis using supervised manifold learning," *Mechanical Systems and Signal Processing*, vol. 23, no. 7, pp. 2301–2311, 2009.

[13] B. Li and Y. Zhang, "Supervised locally linear embedding projection (SLLEP) for machinery fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 25, no. 8, pp. 3125–3134, 2011.

[14] K. Shin, J. K. Hammond, and P. R. White, "Iterative SVD method for noise reduction of low-dimensional chaotic time series," *Mechanical Systems and Signal Processing*, vol. 13, no. 1, pp. 115–124, 1999.

[15] Y.-F. Sang, D. Wang, J.-C. Wu, Q.-P. Zhu, and L. Wang, "Entropy-based wavelet de-noising method for time series analysis," *Entropy*, vol. 11, no. 4, pp. 1123–1147, 2009.

[16] Y.-F. Wang and P. J. Kootsookos, "Modeling of low shaft speed bearing faults for condition monitoring," *Mechanical Systems and Signal Processing*, vol. 12, no. 3, pp. 415–426, 1998.

[17] K. A. Loparo, Bearing Data Center, Case Western Reserve University, 2013, http://csegroups.case.edu/ bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website.

[18] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, Germany, 3rd edition, 2001.

[19] G. Liao, S. Liu, T. Shi, and G. Zhang, "Gearbox condition monitoring using self-organizing feature maps," *Proceedings of the Institution of Mechanical Engineers C: Journal of Mechanical Engineering Science*, vol. 218, no. 1, pp. 119–130, 2004.

[20] H. Qiu, J. Lee, J. Lin, and G. Yu, "Robust performance degradation assessment methods for enhanced rolling element bearing prognostics," *Advanced Engineering Informatics*, vol. 17, no. 3-4, pp. 127–140, 2003.

[21] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Advances in Neural Information Processing Systems*, vol. 17, MIT Press, Cambridge, Mass, USA, 2004.