Chapter 3

Condition Monitoring and Fault Diagnosis of Roller Element Bearing

Tian Ran Lin, Kun Yu and Jiwen Tan

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/67143

Abstract

Rolling element bearings play a crucial role in determining the overall health condition of a rotating machine. An effective condition-monitoring program on bearing operation can improve a machine's operation efficiency, reduce the maintenance/replacement cost, and prolong the useful lifespan of a machine. This chapter presents a general overview of various condition-monitoring and fault diagnosis techniques for rolling element bearings in the current practice and discusses the pros and cons of each technique. The techniques introduced in the chapter include data acquisition techniques, major parameters used for bearing condition monitoring, signal analysis techniques, and bearing fault diagnosis techniques using either statistical features or artificial intelligent tools. Several case studies are also presented in the chapter to exemplify the application of these techniques in the data analysis as well as bearing fault diagnosis and pattern recognition.

Keywords: rolling element bearings, condition monitoring, fault diagnosis

1. Introduction

Rolling element bearings are the most critical but vulnerable mechanical components in a rotating machine. A bearing failure can lead to a complete machine breakdown causing unintended interruption to a production process and financial losses. It is important to have an effective bearing condition monitoring (CM) and fault diagnosis system in place so that incipient bearing faults can be detected and correctly diagnosed on time to prevent them from deteriorating further to cause damage to a machine. For instance, an early detection of incipient defect of a rolling element bearing in a high speed train or a wind turbine can lead to a timely maintenance/replacement to prevent potential disastrous consequence and human loss caused by unexpected failure of critical mechanical components.

Many condition-monitoring and fault diagnosis techniques have been developed in the last few decades to improve the reliability of rolling element bearings. This chapter provides an overview on the most commonly employed condition-monitoring, signal analysis, and fault diagnosis techniques for rolling element bearings and discusses some of the pros and cons of these techniques.

A starting but most fundamental information in bearing condition monitoring is the characteristic bearing defect frequencies. The characteristic defect frequency components in a signal are generated by flaws or faults presented in a bearing when the bearing is operated at a specific machine rotating speed under certain loading conditions. Alternatively, defect signals can also be produced accompanying the normal wear process during a bearing's operational life.

Figure 1 shows the graphical and the cross-sectional representations of a rolling element bearing. The bearing comprises four mechanical components: an outer race, an inner race, rollers (balls), and a cage that hold the rollers (balls) in place. Correspondingly, there are four possible characteristic defect frequencies for a rolling element bearing: ball (roller) pass frequency at the outer race (BPFO), ball (roller) pass frequency at the inner race (BPFI), ball (roller) spin frequency (BSF), and fundamental train frequency (cage frequency) (FTF). The formulae for these four characteristic bearing defect frequencies are listed in **Table 1**.

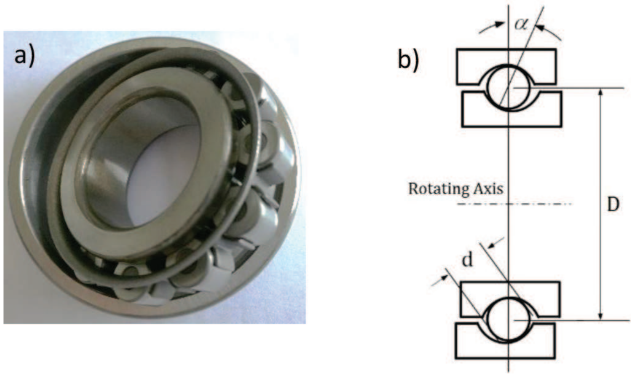

Figure 1. (a) A graphical illustration of a roller element bearing and (b) a cross-sectional view of the roller element bearing.

Ball-pass frequency at outer race (BPFO)	$\mathrm{BPFO} = \frac{n}{2}\frac{N}{60}\left(1-\frac{d}{D}\cos\alpha\right)$
Ball-pass frequency at inner race (BPFI)	$\mathrm{BPFI} = \frac{n}{2}\frac{N}{60}\left(1+\frac{d}{D}\cos\alpha\right)$
Ball-spin frequency (BSF)	$\mathrm{BSF} = \frac{D}{2d}\left[1-\left(\frac{d}{D}\cos\alpha\right)^2\right]$
Fundamental train frequency (FTF)	$\mathrm{FTF} = \frac{1}{2}\frac{N}{60}\left(1-\frac{d}{D}\cos\alpha\right)$

Note: N is the shaft speed in revolutions per minute (RPM), n is the number of roller elements in a bearing, α is the contact angle of the bearing due to the load from the radial plane, d is the diameter of the roller, and D is the mean diameter of the bearing as shown in **Figure 1**.

Table 1. Formulae of the bearing defect frequencies.

A bearing defect signal can be simulated using the following equations [1]:

$$s(t,n) = Qe^{-\gamma(t-\frac{n}{\mathrm{BDF}})}\sin\left[2\pi f_r\left(t-\frac{n}{\mathrm{BDF}}\right)\right]+\frac{O(t)}{r_{\mathrm{sn}}},\ t<\frac{n+1}{\mathrm{BDF}},\ n=0,1,2,\ldots. \tag{1a}$$

and

$$s(t,n) = Qe^{-\gamma(t-\frac{n+1}{\mathrm{BDF}})}\sin\left[2\pi f_r\left(t-\frac{n+1}{\mathrm{BDF}}\right)\right]+\frac{O(t)}{r_{\mathrm{sn}}},\ t\geq\frac{n+1}{\mathrm{BDF}},\ n=0,1,2,\ldots. \tag{1b}$$

where Q is the assumed maximum loading intensity for a bearing defect and t is the time variable, BDF represents a bearing defect frequency, f_r is the assumed bearing resonance frequency and α is the energy decay constant of the bearing race. The first part in Eqs. (1a) and (1b) is the signal produced by a bearing defect, and the second part of the equations is the superimposed white Gaussian noise representing the machine background noise. n is the pulse index of the bearing defect frequency, $O(t)$ is the white Gaussian noise and r_{sn} is the assumed signal-to-noise ratio (SNR). A typical bearing defect signal is shown in **Figure 2**. The parameters used in the simulation of the signal are listed in **Table 2**.

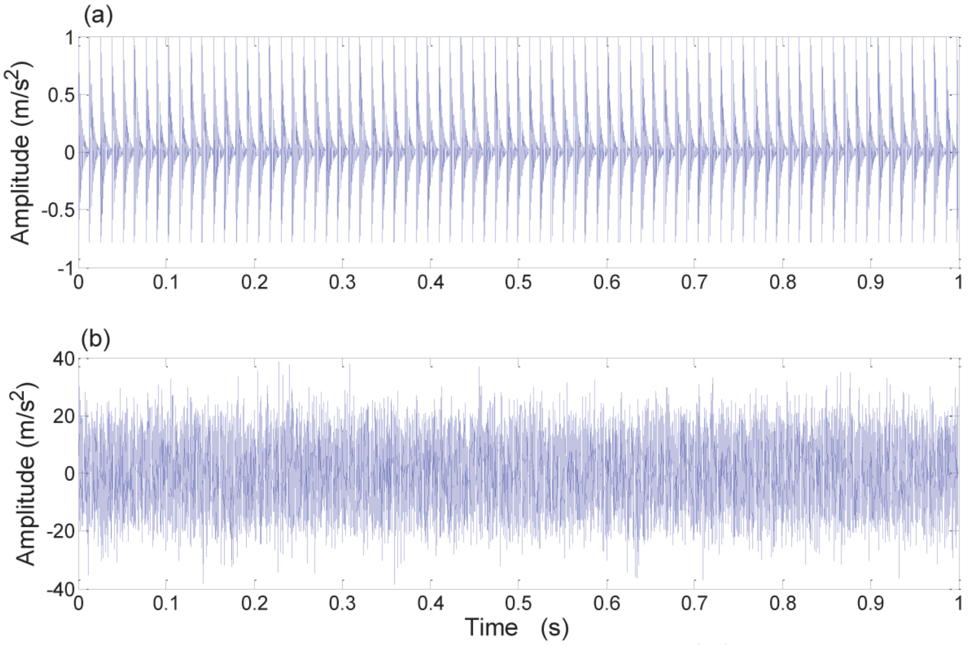

Figure 2. A simulated defect signal of a roller element bearing due to a fault at the outer race: (a) pure defect signal; (b) noise-added signal.

Fault type	Defect frequency (Hz)	Resonance frequency, f_r (Hz)	Decay constant, α	Loading intensity, Q	SNR, r_{sn}
BPFO	78.18	4000	200	1	−20 dB

Table 2. Parameters used in the simulation.

2. Condition-monitoring techniques

Condition-monitoring (CM) techniques are required to acquire the operation condition data of a rotating machine. The CM data can then be processed and analyzed using appropriate signal analysis techniques to obtain the most relevant characteristic parameters before being used in a diagnosis or prognosis algorithm to evaluate/predict the health state of the machine. The data employed in a condition-monitoring program can be vibration, noise, electric current, oil and grease, temperature or a combination of these data. The CM data can be analyzed using either time-domain techniques, frequency-domain techniques, or time-frequency techniques according to the data properties such as linearity or nonlinearity, stationary or nonstationary, to extract the most useful and effective features for bearing fault diagnosis. This section summarizes some of the most commonly employed techniques in condition-monitoring applications of rolling element bearings.

2.1. Vibration technique

Vibration technique is the most frequently employed technique for machine condition monitoring. A change of the vibration signal in a machine without changing the operation condition can imply a change of health state of the machine. Vibration signals are typically generated by defects in the moving components of the machine such as defects in a bearing, gearboxes, reciprocating components, and so on. For example, when a rolling element bearing operates under a faulty condition, an impulse signal will be generated as other bearing components passing through the faulty position. This will lead to an increase in overall vibration amplitude of the machine. The defect component of the bearing and its severity can be determined by the characteristic defect frequency component contained in the condition monitoring signal (see **Figure 2**). The frequency range of vibration measurement can be as low as in the infrasonic region (below 20 Hz) and span across to the upper limit of the audible frequency (20 kHz). Though the vibration technique can be problematic in acquiring useful signals when it is deployed for condition monitoring of large low-speed rotating machine where the energy of an incipient defective signal is usually weak and often submerges under the background noise. High-frequency techniques such as acoustic emission (AE) technique can be employed to overcome this limitation.

2.2. Acoustic emission technique

Acoustic emission is a transient elastic wave generated by the sudden release or redistribution of stress in a material. For example, in bearing condition-monitoring applications, AE is generated by the sudden release of energy caused by the material deformation as other bearing components passing through a defect part. The signal then propagates to the bearing house to be detected by a monitoring AE sensor. Compared to vibration signals, the AE signal is less likely to be affected by the dominating noise and vibration generated by the moving mechanical components of the monitored system due to the high-frequency nature of the signal (typically above 100 kHz). Care should be taken in choosing the mounting locations of AE sensors to minimize the energy loss along the AE propagation path for better signal clarity. Furthermore, AE also comes with inherited problems such as calibration, nonlinearity, data storage and transfer, data processing and interpretation. Recently, Lin et al. [1–3] developed a number of signal-processing algorithms to overcome the problems of nonlinearity and large AE data. Nowadays, a large industry deployment of AE technique in bearing condition monitoring is still restricted by the expensive highly specialized AE data acquisition devices.

2.3. Current signal technique

Current signal technique is mainly used to monitor the bearing condition of electric motors. The technique is based on the principle that the vibration signal generated in a motor is closely related to the change of magnetic flux density in the motor. Stator current technique is the frequently employed current signal technique. When a motor bearing operates on a faulty condition, the radial motion of the motor axis would lead to a small shift of the rotor causing a change in the magnetic flux density between the stator and the rotor. The induced voltage will then cause the variation of the stator current. The stator current technique employs noninvasive sensors to monitor the variation of the stator current; therefore, it is easy to implement and simple to operate.

2.4. Oil and debris-monitoring technique

The application of oil and debris-monitoring technique is a frequently employed tribology approach for machine condition monitoring. The health condition of a roller element bearing can be monitored by analysis of the properties and particle contents of the lubrication oil. The application of oil and debris-monitoring technique is rather straightforward. Though such tribology analysis is normally undertaken at laboratories using spectrometers and scanning electron microscopes rather than *in situ*. The technique is also limited to condition-monitoring applications of lubrication-related or wear-related problems.

2.5. Thermography technique

Thermography technique detects faults in a bearing by measuring the emission of infrared energy using thermo-infrared devices during bearing operation process. Laser thermography devices or thermo-infrared cameras are the most common instruments used for such measurement. This technique can be applied to monitor the heat variation caused by the change of bearing lubrication, load, and operation speed. It is particularly sensitive in monitoring the overheating phenomenon caused by improper lubrication but less sensitive in detecting incipient fault developed in a bearing such as early pitting, slight wear, or peeling.

3. Data analysis and fault diagnosis techniques

Many signal analysis and machine fault diagnosis techniques have been developed in the last few decades in order to improve the reliability, efficiency, and lifespan of machines, as well as to reduce the maintenance and operation cost. In this section, the most frequently employed signal-processing and fault diagnosis techniques for rolling element bearings are briefly discussed and summarized.

3.1. Time-domain data analysis techniques

3.1.1. Time series analysis

Time series analysis is a mathematical method, which handles an observed data series in a statistic manner. Time series analysis is based on the assumption that the variation trend of a historical observed data can be extended to provide an indication/prediction of future data variation of a same monitored system. A typical time series analysis approach is to establish a predictive model based on the observed data series. The three widely adopted univariate time series models in machine fault diagnosis are autoregressive (AR) model, moving average (MA) model, and autoregressive moving average (ARMA) model [4].

A general linear ARMA (p, q) model can be expressed as

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{2}$$

where y_t is the time series needed to be modeled, c is a constant, p is the number of autoregressive orders, q is the number of moving average orders, ϕ_i is the autogressive coefficients, θ_j is the moving average coefficients, and ε_t is the independent and identically distributed terms, which are commonly assumed to have zero mean and a constant variance, or (i) $E(\varepsilon_t) = 0$; (ii) $E(\varepsilon_t \varepsilon_T) = 0$ for t ≠ T ; and (iii) $E(\varepsilon_t^2) = \sigma^2$.

An AR model and an MA model can be viewed as the special case of an ARMA model. For instance, when all autoregressive coefficients ϕ_i equal to 0 $(\phi_i = 0,\ 1 \le i \le p)$, an ARMA model degrades to an MA model. On the other hand, when all moving average coefficients θ_j equal to 0 $(\theta_j = 0,\ 1 \le j \le q)$, an ARMA model degrades to an AR model.

After selecting the most suitable time series model for the data analysis, the next step is to determine the orders of AR or MA models. The commonly adopted order parameter determination criteria for these time series models are the final prediction error (FPE) criterion, Akaike information criterion (AIC), and Bayesian information criterion (BIC).

On the other hand, the common methods used to determine the autoregressive coefficients and moving average coefficients are least squares estimation, maximum likelihood estimation, Yuler-Walker estimation, and so on.

It is worth mentioning that ARMA models are developed based on the assumption that the signal is stationary. If the signal is not stationary, some data preprocessing steps need to be adopted. For example, (1) performing a differential operation of the data term by term until the signal satisfies the stationary criterion and (2) decomposing a nonstationary signal by empirical mode decomposition (EMD) to obtain the stationary intrinsic mode functions (IMFs).

3.1.2. Minimum entropy deconvolution

Minimum entropy deconvolution (MED) technique is proposed originally by Wiggins, which has been successfully employed in dealing with the seismic response signal [5]. The basic idea of MED is to find an inverse filter that counteracts the effect of the transmission path [6]. It is designed to reduce the spread of impulse response functions (IRFs) and then obtain signals, which are close to the original impulses giving rise to them. The MED technique has been successfully employed in bearing fault diagnosis such as in [6].

Figure 3 illustrates the basic idea of the MED technique. In this process, an unknown impact signal $x(n)$, which can be as highly impulsive as possible, passes through a structural filter **h** and then mixes with a noise $e(n)$ to produce an intermediate output $o(n)$. The signal $o(n)$ then passes through an inverse (MED) filter **f** to produce a final output $y(n)$. The process eliminates the structure resonance and the final output $y(n)$ after the inverse filter needs to be as close as possible to the original input $x(n)$.

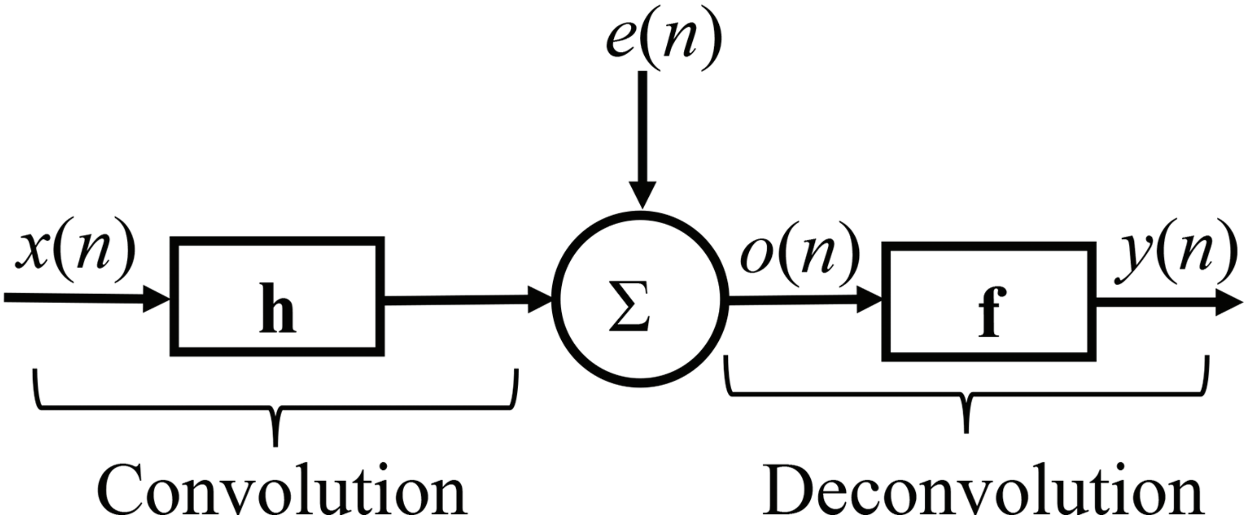

Figure 3. Inverse filtering (deconvolution) process for MED.

The inverse filter *f(l)* can be modeled as a finite impulse response (FIR) filter with *L* coefficients:

$$y(n) = \sum_{l=1}^{L} f(l)x(n-l), \tag{3}$$

and

$$f(l) * h(l) = \delta(n - l_m). \tag{4}$$

where δ is a Dirac delta function and l_m is the time delay after the MED process, which displaces the entire signal by l_m but keeps the impulse spacing of the signal.

The inverse filter *f(l)* is implemented for the MED technique by the objective function method (OFM). The OFM is an optimization process designed to maximize the kurtosis of the output signal, $y(t)$. OFM achieves this by changing the coefficients of the filter *f(l)*. The kurtosis is taken as the normalized fourth-order moment given by

$$K\big(f(l)\big) = \sum_{n=1}^{N} y^4(n) / \left[\sum\nolimits_{n=1}^{N} y^2(n)\right]^2, \tag{5}$$

and the maximum kurtosis of *y(t)* can be obtained according to *f*(*l*) for which the derivative of the objective function is zero:

$$\partial K(f(l))/\partial f(l) \ = 0 \tag{6}$$

where the filter coefficients of *f(l)* can converge to a given tolerance by the iterative process [7].

3.1.3. Spectral kurtosis

Spectral kurtosis (SK) was first proposed in the 1980s for detecting impulsive events in sonar signals. SK was first applied in bearing fault diagnosis by Antoni [8]. The method basically computes a kurtosis at "each frequency line" in order to discover the presence of hidden nonstationarities and to indicate in which frequency band it takes place. The method has been proved to be relatively robust against strong additive noise. An SK of nonstationary signals is defined based on the Wold-Cramer decomposition, which describes any stochastic random process *x(t)* as the output of a causal, linear, and time-varying system [9]:

$$x(t) = \int_{-1/2}^{+1/2} H(t,f) \mathrm{e}^{\mathrm{j}2\pi f n} \mathrm{d}Z_x(f), \tag{7}$$

where $\mathrm{d}Z_x(f)$ is an orthonormal spectral process of unit variance and $H(t, f)$ is the time-varying transfer function interpreted as the complex envelope of $x(t)$ at frequency f. The SK of a signal x (t) is defined as a normalized fourth-order spectral accumulation given by [8, 9]:

$$K_x(f) = \frac{\langle |H(t,f)|^4 \rangle}{\langle |H(t,f)|^2 \rangle^2} - 2, \tag{8}$$

in which $K_x(f)$ is the spectral kurtosis of signal $x(t)$ around frequency f and $\langle \cdot \rangle$ denotes the averaging over time.

Antoni and Randall [9] proposed two techniques to calculate the spectral kurtosis, one is based on short-time Fourier transform (STFT) (the so-called kurtogram for finding the optimal filter) and the other is based on one-third binary filter banks (fast kurtogram for on-line condition monitoring and fault diagnosis). Kurtogram is a powerful tool for the analysis of nonstationary signals in bearing fault diagnosis, though it has been reported that the technique fails to detect a bearing fault when the defect signal has low signal-to-noise ratio and contains non-Gaussian noise with high peaks [10, 11].

3.1.4. Singular-value decomposition

Singular-value decomposition (SVD) is a numerical method which states that a matrix **A** of rank L can be decomposed into the product of three matrices: **U** (an orthogonal matrix), **Λ** (a diagonal matrix) and $\mathbf{V}^{\mathrm{T}}$ (the transpose of an orthogonal matrix **V**) [12]. This method is usually presented as

$$\mathbf{A}_{m\times n} = \mathbf{U}_{m\times m} \cdot \mathbf{\Lambda}_{m\times n} \cdot \mathbf{V}^{\mathrm{T}}_{n\times n}, \tag{9}$$

where $\mathbf{U}^{\mathbf{T}}\mathbf{U} = 1$ and $\mathbf{V}^{\mathbf{T}}\mathbf{V} = 1$; $\mathbf{\Lambda}_{m\times n}$ is a diagonal matrix containing the square roots of eigenvalues of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ which can be expressed as $\mathbf{\Lambda} = \mathrm{diag}(\sigma_1,\ \sigma_2,\ \ldots, \sigma_L)$, where $L = \min(m, n)$.

The non-zero diagonal terms $\sigma_i(i = 1, 2, \cdots, l)$ together with the zero terms $\sigma_{l+1} = \ldots = \sigma_L = 0$ form the singular values of the matrix **A**.

SVD can be an effective tool for data reduction, for example, a reduction of the fault feature dimensions. It is also a useful tool for signal de-noising. The application of SVD in signal de-noising is mainly operated as follows:

Suppose a bearing CM signal $x(k)$ can be modeled as

$$x(k) = y(k) + n(k), \tag{10}$$

where $y(k)$ and $n(k)$ are respectively, the uncontaminated signal and noise. A conversion method such as phase space reconstruction can be used to transform the signal from a one-dimensional vector into a two-dimensional matrix as follows:

$$\mathbf{A} = \overline{\mathbf{A}} + \mathbf{N}, \tag{11}$$

where the matrix $\overline{\mathbf{A}}$ represents the uncontaminated data $y(k)$, which contains characteristic fault information and the matrix **N** signifies the unwanted noise part $n(k)$. From Eqs. (9) and (11), we have

$$\mathbf{A} = \overline{\mathbf{A}} + \mathbf{N} = [\,\mathbf{U}_l \quad \mathbf{U}_0\,]\begin{bmatrix} \mathbf{\Lambda}_l & 0 \\ 0 & \mathbf{\Lambda}_0 \end{bmatrix}\begin{bmatrix} \mathbf{V}_1^{\mathrm{T}} \\ \mathbf{V}_0^{\mathrm{T}} \end{bmatrix}. \tag{12}$$

$\mathbf{\Lambda}_1$ in Eq. (12) contains the significant singular values $\sigma_i(i = 1, 2, \cdots, l)$ which are used to construct the uncontaminated data and $\mathbf{\Lambda}_0$ contains small singular values σ_i $(i = l + 1, \ldots, L)$, which can be viewed as noise.

From Eq. (12), the de-noising signal matrix can be written as

$$\overline{\mathbf{A}} = \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{V}_1^{\mathrm{T}} \tag{13}$$

$\overline{\mathbf{A}}$ is a reconstructed matrix using only the largest l number of singular values whose values are greater than a pre-set threshold value ε. The rest of the singular values are replaced by zero such that

$$\sigma_i = 0 \text{ when } \sigma_i \leq \varepsilon,\ t = l + 1, \ldots, L. \tag{14}$$

3.2. Frequency-domain data analysis techniques

3.2.1. Power spectrum

Spectrum analysis is the most popular frequency-domain analysis techniques, which transforms a time-domain data into discrete frequency components by Fourier transform. The power spectrum of a time-domain signal is defined as the square of the magnitude of the Fourier transform of a signal. It can be written as

$$P(\omega) = |\int_{-\infty}^{+\infty} x(t)\mathrm{e}^{-\mathrm{j}\omega t} dt|^2 = X(\omega)X^*(\omega), \tag{15}$$

where $X(\omega) = \int_{-\infty}^{+\infty} x(t)e^{-\mathrm{j}\omega t} dt$ is the Fourier transform of a signal, $X^*(\omega)$ is its complex conjugate and ω is the angular frequency in radian/s. The power spectrum is frequently employed to extract useful characteristic defect frequency components of a stationary CM signal.

3.2.2. Cepstrum

A cepstrum is defined as the power spectrum of the logarithm of the power spectrum of a signal [13]. The name of cepstrum was derived by reversing the first four letters of spectrum. There are four types of cepstra: a real cepstrum, a complex cepstrum, a power cepstrum and a phase cepstrum.

The real cepstrum of a signal $x(t)$ is given by

$$\mathrm{c}(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log|X(\omega)|\mathrm{e}^{\mathrm{j}\omega t} d\omega. \tag{16}$$

The complex cepstrum of a signal $x(t)$ is given by

$$\vec{c}\,(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log\Big(X(\omega)\Big)\mathrm{e}^{\mathrm{j}\omega t} d\omega. \tag{17}$$

The power cepstrum of a signal $x(t)$ is given by

$$c(t)^2 = \frac{1}{2\pi} |\int_{-\pi}^{\pi} \log|X(\omega)|\mathrm{e}^{\mathrm{j}\omega t} d\omega|^2. \tag{18}$$

The most commonly used cepstrum in machine condition monitoring is power cepstrum. The following steps can be taken to calculate the power cepstrum of a signal: A signal → power spectrum of the signal → logarithm of the power spectrum → power spectrum of the data from the previous step → inverse Fourier transform of the log power spectrum from the previous step (power cepstrum). The main application of cepstrum analysis in machine condition monitoring is for signals containing families of harmonics and sidebands where it is the whole family rather than individual frequency component characterizing the fault [13] (typical for bearing CM signals). The technical terms used in a cepstrum are quefrency, gamnitude and rahmonics corresponding to frequency, amplitude and harmonics in a spectrum analysis. The cepstrum can be used for the harmonics generated by bearing faults, but only if they are well separated. In contrast, envelope analysis to be introduced in the next section is not limited by such restriction.

3.2.3. Envelope spectrum

It has been pointed out by Randall [13] that the benchmark method for rolling element bearing diagnosis is envelope analysis as the spectrum of raw bearing CM signals often contains little information about bearing faults. In envelope analysis, a time waveform is bandpass filtered in a high-frequency band and the fault signal is amplified by the structural resonances and amplitude modulated to form the envelope signal for bearing diagnosis [13]. The procedures for envelope analysis are briefly described below:

Given a real signal $x(t)$, its Hilbert transform, $h(t) = H\{x(t)\}$ is defined as

$$h(t) = H\{x(t)\} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d\tau = \frac{1}{\pi\tau} x(t), \tag{19}$$

The real signal $x(t)$ and its Hilbert transform $h(t)$ together form a new complex analytic signal $\mathrm{z}(t)$,

$$\mathrm{z}(t) = x(t) + jh(t). \tag{20}$$

The envelope signal $E(t)$ is simply the absolute value of the analytic signal $\mathrm{z}(t)$,

$$E(t) = |z(t)| = |x(t) + jh(t)| = \sqrt{x^2(t) + h^2(t)}. \tag{21}$$

After taking a fast Fourier transform on the envelope signal $E(t)$, an envelope spectrum can be obtained. An envelope spectrum can reveal the repetition characteristic defect frequencies caused by bearing faults even in the presence of a small random fluctuation. The envelope spectrum of the noise-added bearing defect signal shown in **Figure 2(b)** is given in **Figure 4**. It is noted that due to a high noise level in the signal (SNR = 0.05, representing an initial bearing defect), the envelope spectrum is compromised by many artificial frequency components and produces a subtle bearing defect information. Also due to such interference, the calculated defect frequency (the modulated frequency) of the outer race fault and its higher harmonics (1 × BPFO, 2 × BPFO and so on) shown in the envelope spectrum are also slightly lower than that of the simulated defect frequency (1 × BPFO = 78.18 Hz).

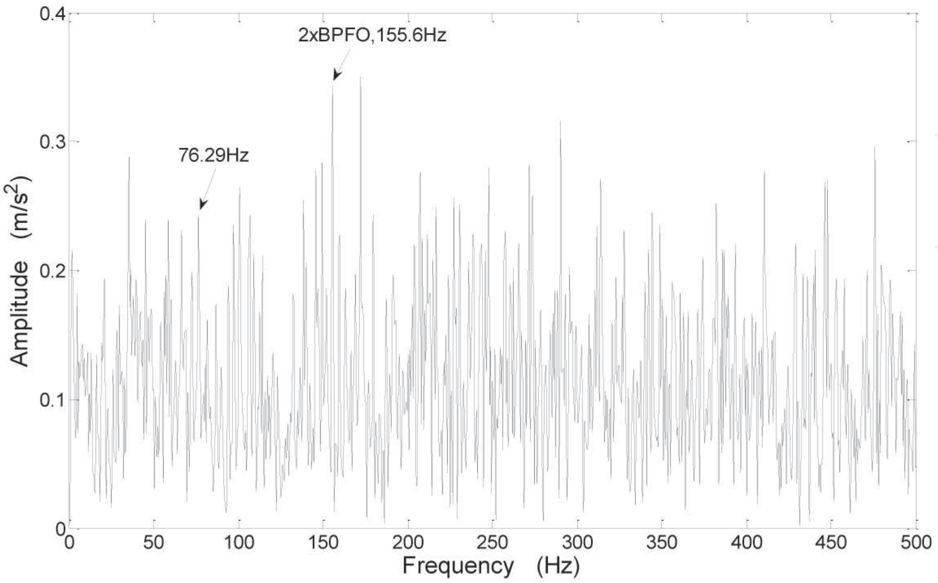

Figure 4. Envelope spectrum of the noise-added bearing defect signal shown in **Figure 2(b)**.

3.2.4. Higher-order spectra

Higher-order spectra (HOS) (also known as polyspectra) consist of higher-order moment of spectra, which are able to detect nonlinear interactions between frequency components [14]. Higher-order spectra are defined as the Fourier transform of the corresponding cumulant sequences of a signal. For instance, the first-order cumulant of a stationary process is the mean of the signal:

$$C_1x = E\{x(n)\}. \tag{22}$$

The second- and third-order cumulants of a stationary process are defined as

$$C_2x(k) = E\{x^*(n)x(n+k)\}, \tag{23}$$

and

$$C_3x(k,l) = E\{x^*(n)x(n+k)x(n+l)\}–C_2x(k)C_2x(l-m)–C_2x(l)C_2x(k-m) \tag{24}$$

From Eqs. (23) and (24), it is clear that the power spectrum is, in fact, the FT of the second-order cumulant of a signal and the third-order spectrum also termed as bispectrum is the FT of the third-order cumulant. Note that the bispectrum $S_{3x}(\omega_1, \omega_2)$ is a function of two frequencies. Therefore, it can detect phase coupling between two frequency components which appears as a third frequency component at the sum or difference of the first two (frequencies) with a phase that is also the sum or difference of the first two.

Traditionally, power spectrum is used to break down a time waveform signal into a series of frequency components. However, power spectrum cannot determine whether peaks at harmonically related positions are phase coupling since power spectrum uses only the magnitude of Fourier components and the phase information is neglected. Higher-order spectra such as bispectrum use the phase information of the signal and are capable to detect phase coupling of frequency components in the spectra. Therefore, a bispectrum can provide additional phase information than a power spectrum analysis.

The motivation behind the use of higher-order spectrum analysis is summarized as follows. Firstly, the technique can suppress Gaussian noise in the data processing of unknown spectral characteristics for fault detection, parameter estimation and classification problems. If Gaussian noise is embedded in a non-Gaussian signal, a HOS transform can eliminate the noise. On the other hand, periodic, quasi-periodic signals and self-emitting signals from complex machinery in practical applications are typical non-Gaussian signals which will be preserved in the transform. Secondly, a HOS analysis can preserve the phase information. For example, there are situations in practice in which the interaction between two harmonic components creates a third component at their sum and/or difference frequencies. Thirdly, HOS can play a key role in detecting and characterizing the type of nonlinearity in a system from its output data.

For a better illustration of the signal-processing techniques discussed above in data analysis, a number of case studies are given in the following section to exemplify the usage of the above algorithms in bearing fault detection.

3.2.5. Case studies

Case 1: (AR & MED de-noising): In this case study, the AR model described in Section 3.1.1 and the MED method described in Section 3.1.2 are employed in the analysis to filter the noise-added bearing defect signal shown in **Figure 2(b)** prior to an envelope analysis to enhance the bearing defect frequency components in the envelope spectrum for a more reliable bearing fault diagnosis. The results are presented in **Figures 5** and **6**. Compared to the result in **Figure 4**, it is shown that the two-step de-noising by the AR and MED models has successfully suppressed the artificial components and enhanced the defect signal representation in the spectrum.

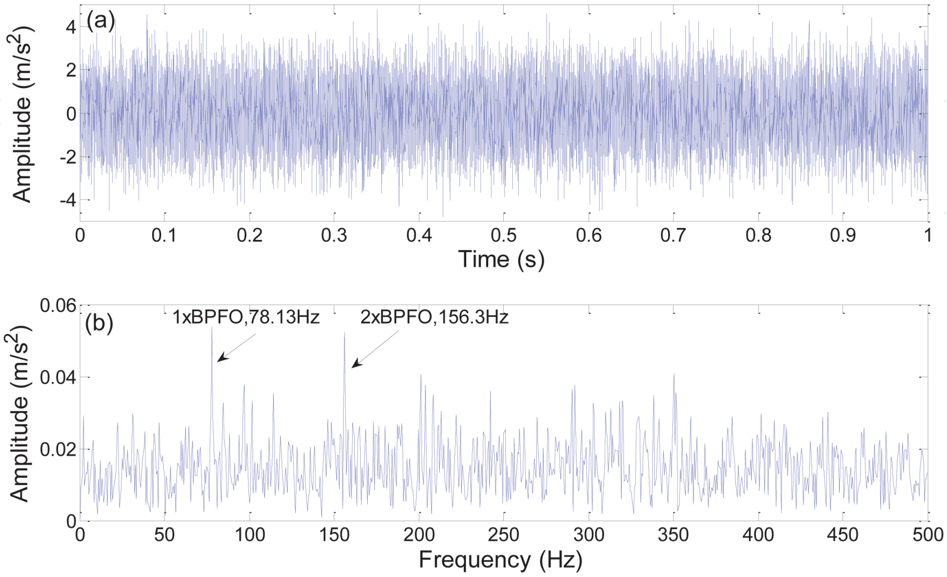

Figure 5. The time waveform and the envelope spectrum of the simulated bearing defect signal after preprocessed by an AR model: (a) the time waveform after filtered by an AR model; (b) envelope spectrum of the signal shown in **Figure 5(a)**.

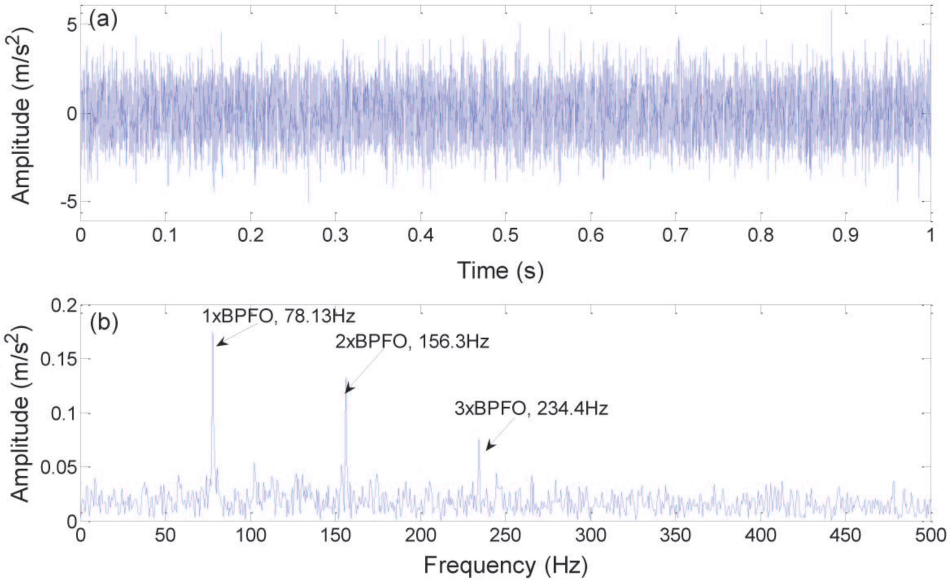

Figure 6. The time waveform and the envelope spectrum of the simulated bearing defect signal after preprocessed by an AR model and filtered further by an MED model: (a) the time waveform after de-noising; (b) envelope spectrum of the signal shown in **Figure 6(a)**.

Case 2 (spectrum kurtogram): The signal used in this case study is shown in **Figure 7**. The signal is generated by adding 0-dB white noise to the simulated bearing defect signal presented in **Figure 2(a)**. In the analysis of the signal, the fast kurtogram algorithm [9] described in Section 3.1.3 is first employed to determine the bearing resonance band (to obtain the center frequency and the bandwidth) having the highest band energy (corresponding to the highest kurtosis value) in the signal. A five-level fast kurtogram based on a filter band and fast Fourier transform is shown in **Figure 8** and the highest band energy is found to occur in the band encircled by the white ellipse in the figure. The band is found to be centered at 3958.33 Hz (close to the simulated bearing resonance frequency of 4000 Hz as listed in **Table 2**) and has the bandwidth of 416.67 Hz. The band has the highest kurtosis value of 0.1 which occurs at level 4.5 in the decomposition. The next step after the decomposition is to take an envelope analysis based on the band-filtered optimum frequency band signal obtained from the fast kurtogram and the result is shown in **Figure 9(a)** and **(b)**. It is shown that the spectrum kurtosis technique can detect the characteristic defect frequency from weak-bearing defect signals.

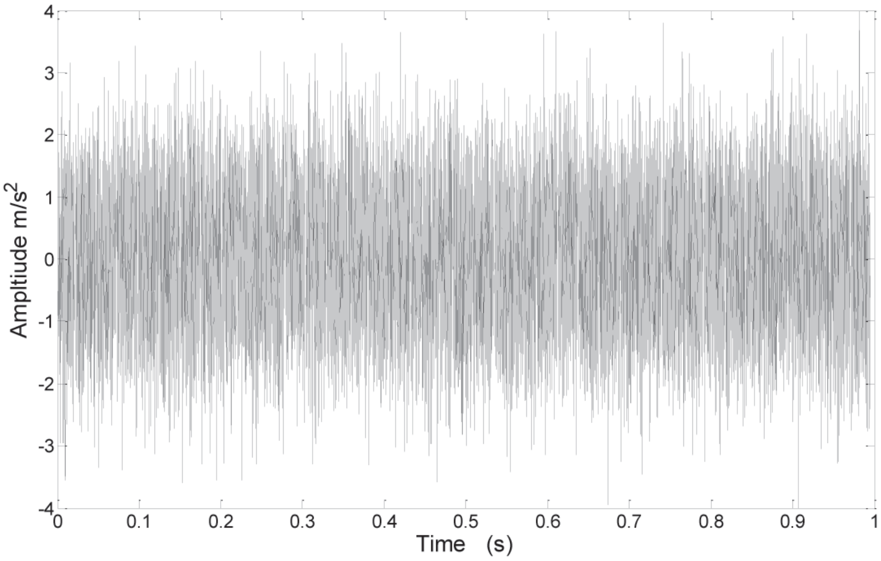

Figure 7. Simulated bearing defect signal with 0-dB white noise added.

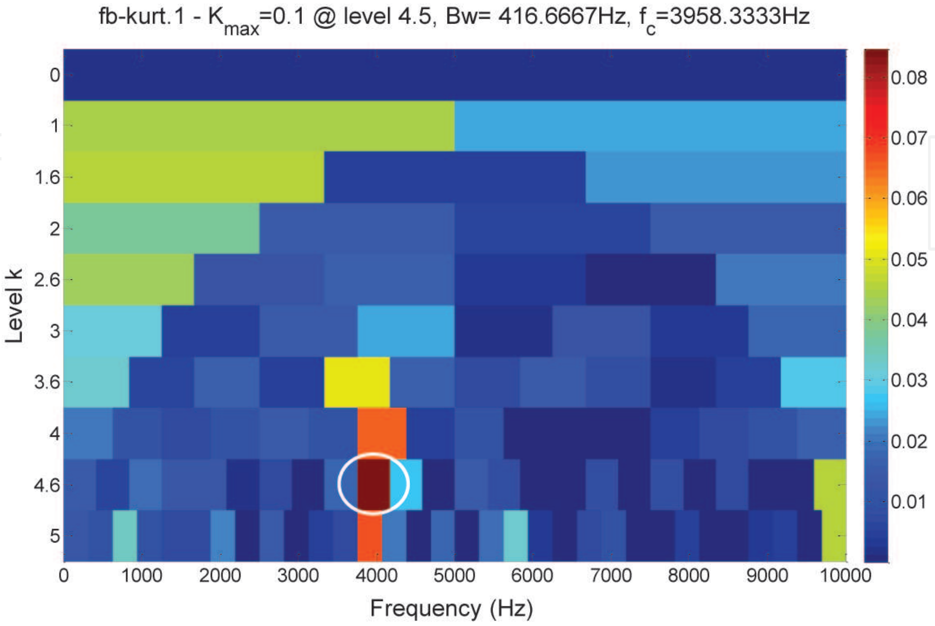

Figure 8. Fast kurtogram calculated based on the defect signal shown in **Figure 7**.

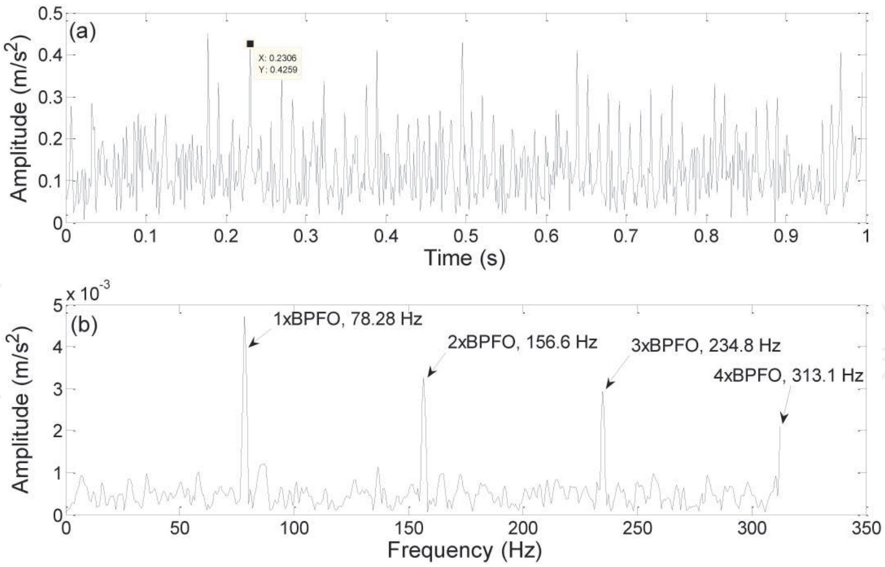

Figure 9. Spectrum kurtosis analysis of the bearing defect signal: (a) envelope of the band-filtered signal; (b) Fourier transform of the envelope signal.

3.3. Time-frequency analysis

A major limitation of frequency-domain analysis is that it is only useful in dealing with stationary signals. The presence of transient or nonstationary signals in the data would not be captured in a traditional frequency-domain analysis. To overcome this limitation, time-frequency analysis techniques are then developed for a better understanding of how spectrum properties change with time. In time-frequency analysis, waveform signals are analyzed in both time and frequency domains to capture the progressive change of spectrum components. The most commonly employed time-frequency analysis techniques are short-time Fourier transform (STFT), wavelet transform (WT), Wigner-Ville distribution (WVD) and adaptive signal analysis techniques such as empirical mode decomposition (EMD) technique.

3.3.1. Short-time Fourier transform

Short-time Fourier transform adds a time variable to the traditional Fourier spectrum, thus allowing it to investigate the time-varying nature of a signal. In a SFFT analysis, a continue time-domain waveform is multiplied by a sliding narrow time window and a Fourier transform is computed on the windowed signal at each time step of the sliding time window. The STFT analysis is based on the consideration that if a signal can be considered stationary over the length of the chosen sliding time window, a Fourier transform can be performed on the windowed signal segment for each new position of the sliding time window to obtain a satisfactory time-frequency analysis of a nonstationary signal.

A short-time Fourier transform of a continuous signal $x(t)$ can be computed by

$$\mathrm{STFT}_x(\tau,\omega) = \int_{-\infty}^{+\infty} x(\tau) w^*(\tau - t) \mathrm{e}^{-\mathrm{j}\omega\tau} d\tau, \tag{25}$$

where $w(\tau - t)$ is a finite time window function centered at time t. The asterisk sign (*) in the time window indicates a complex conjugate and the analysis window can be regarded as the impulse response of a low-pass filter. A spectrogram, which is the squared amplitude of an

STFT transform, is often used in the display of the transformation result for signal analysis and interpretation. A major limitation of STFT is that the analysis has a uniform resolution in both time and frequency planes implying that a small time window will have a good time resolution but poor frequency resolution and vice versa. This drawback limits the technique for the analysis of signals with slow-changing dynamic only. Another time-frequency analysis technique, wavelet transform, has then been developed to overcome this limitation.

3.3.2. Wavelet transform

Wavelet transform (WT) is a time scale representation of a signal. It is the inner product of a signal with the translated and scaled family of a mother wavelet function *ψ(t)*. In general, WT analysis can be categorized into three forms: a continuous wavelet transform (CWT), a discrete wavelet transform (DWT), and a more general wavelet packet decomposition (WPD).

In CWT, the wavelet transform of a continuous signal $x(t)$ is calculated using

$$W_f(u,s) = \frac{1}{\sqrt{s}} \int x(t)\psi^* \left(\frac{t-u}{s}\right) dt. \tag{26}$$

where s is the scale parameter and u is the time translation. If one considers the wavelet function *ψ(t)* as a bandpass impulse response, then the wavelet transform is simply a bandpass analysis. Care should be taken to ensure that the decomposed signal can be perfectly reconstructed from its wavelet representation when using a wavelet transform. Thus, a WT has to meet the criterion which is also known as the admissibility condition.

A CWT is mainly used in data analysis of scientific research. In practical applications, the discrete version, a discrete wavelet transform (DWT), is more popular due to the small computation cost and excellent signal compaction properties. A DWT decomposes a signal into different frequency bands by passing it through a series of filters. In each step of decomposition, a signal is passed through a pair of low- and high-pass filters simultaneously accompanying by down sampling. A more general form of wavelet analysis is the so-called wavelet packet decomposition (WPD). In WPD, a signal is split into two parts, one contains a vector of approximation coefficients and the other contains a vector of detail coefficients in each stage of decomposition. Both the details and the approximations can be split further in the next level of decomposition which offer a great range of possibilities to decode a signal than ordinary wavelet analysis. Wavelet transform has been widely employed in signal processing for condition monitoring and fault diagnosis of rotating machine [15].

3.3.3. Wigner-Ville distribution

Another popular time-frequency analysis technique is Wigner-Ville distribution (WVD). WVD is the core distribution for the quadratic class of quadratic time-frequency distributions. It yields an ideal resolution for mono-component, linearly frequency-modulated signals, but produces undesired cross-terms for multicomponent and nonlinearly frequency-modulated signals. Wigner-Ville distribution can be viewed as a particular case of the Cohen class distributions which yields a time-frequency energy density computed by correlating the signal with a time and frequency translation of itself.

The WVD of a signal $x(t)$ is defined by

$$W_x(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} x\left(t+\frac{\tau}{2}\right)\cdot x^*\left(t-\frac{\tau}{2}\right)\cdot e^{-j\omega\tau}d\tau, \tag{27}$$

where x^* denotes the conjugate of x. Thus, the Wigner-Ville integral is in fact a Fourier transform of the inner product of a signal and its conjugate with a time delay variable τ. The bilinear nature of this procedure therefore avoids the loss of time-frequency resolution in the transform such as the resolution problem encountered when performing the finite sliding time windowing in STFT.

Compared with other distributions, the WVD has the desirable property of fulfilling the marginal condition, thus the total signal energy can be calculated in time or in frequency using the Plancherel formula:

$$||x^2|| = \int_{-\infty}^{+\infty}|x(t)|^2\mathrm{dt} = \frac{1}{2\pi}\int_{-\infty}^{+\infty}|X(\omega)|^2 d\omega \tag{28}$$

The values $||x(t)||^2$ and $||X(\omega)||^2$ can be interpreted as the energy densities in time and frequency domain, respectively. This enables a direct computation of the energy present at a given time-frequency box from the WVD output.

Another important feature of WVD is that its first conditional moment for a given time t_c equals the instantaneous frequency:

$$\frac{d\varphi}{dt} = \langle\omega\rangle_{t_c} = \frac{1}{|s(t_c)|^2}\int_{-\infty}^{+\infty}\omega W(t_c,\omega)d\omega. \tag{29}$$

Therefore, it can be computed as the average of all frequencies ω present in the time-frequency plane at a time t_c.

In discrete form, the WVD is defined as

$$\mathrm{WVD}(n,k) = \sum_{p=-N}^{N-1} R[n,q]\cdot e^{-j2\pi kq/N}, \tag{30}$$

where $R[n, q]$ is the instantaneous correlation given by

$$R[n,q] = x\left[n+\frac{q}{2}\right]\cdot x^*\left[n-\frac{q}{2}\right], \tag{31}$$

in which n is the number of samples of the analytical or interpolated form of the discrete signal $x[n]$ and q is an odd integer.

Since the instantaneous correlation is centered on a value, the delay q is distributed between the delayed sample $x[n-q/2]$ and the corresponding advanced sample, $[n+q/2]$. It is thus necessary to calculate the value of $x[n]$ at the two half integer positions using an interpolation. In addition, positions at the extremes of $x[n]$ are padded with zeros in order to compute the Fourier transform. A major drawback of WVD analysis is that it can induce artifacts and negative values which need to be properly compensated in the signal analysis [16].

3.3.4. Adaptive signal decomposition

Empirical mode decomposition (EMD) is an adaptive time-frequency analysis technique originally proposed by Huang [17] in 1998. It is based on the local characteristic time scales of a signal and can decompose the signal into a set of complete and almost orthogonal components termed as intrinsic mode functions (IMF). Lei et al. [18] provided a review on the successful application examples of EMD technique in fault diagnosis of rotating machines.

In EMD analysis, the decomposed signal can be represented by

$$\mathrm{x}(t) = \sum_{i=1}^{N} C_i(t) + r_N(t), \tag{32}$$

where $C_i(t)$ represents the ith IMF component and $r_N(t)$ is the residual after the EMD decomposition.

The IMFs represent the natural oscillatory modes imbedded in the signal and can serve as the basis functions, which are determined by the signal itself, rather than predetermined kernels. Thus, the decomposition is a self-adaptive signal process suitable for nonlinear and nonstationary data analysis.

Although the EMD technique has been successfully employed in the analysis of nonlinear and nonstationary signals in various applications, the algorithm itself also has a number of weaknesses, for instance, a lack of a solid theoretical foundation, end effects, a sifting stop criterion and extremum interpolation. To overcome some of these deficiencies, an improved EMD algorithm or a so-called ensemble empirical mode decomposition (EEMD) technique has been developed [19]. EEMD is a noise-assisted data analysis technique which imposes a white Gaussian noise into a signal and then decomposes the mixed signal by using the EMD algorithm. A major advantage of the EEMD technique is that no missing scales will be presented in the decomposition and the IMF components in different scales of the signal are automatically projected into proper reference scales established by the white noise in the background [20].

3.3.5. Case studies

Case 3 (EMD de-noising): In this analysis, the noise-added signal shown in **Figure 7** is adaptively decomposed into 14 IMF components and a residual component using the EMD technique described in the previous section. The decomposition result is shown in **Figure 10**. The correlation coefficient of each IMF component with the original signal can be calculated using

$$\rho_{xy} = \frac{n\sum x_i y_{j,i} - \sum x_i \sum y_{j,i}}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_{j,i}^2 - \left(\sum y_{j,i}\right)^2}}, \tag{33}$$

where x_i $(i = 1, 2, \ldots, n)$ is the original data, $y_{j,i}$ $(i = 1, 2, \ldots, n)$ is the data of one of the jth IMF components and n is the data record length. The calculated correlation coefficients for the 14 IMF components and the residual are listed in **Table 3**.

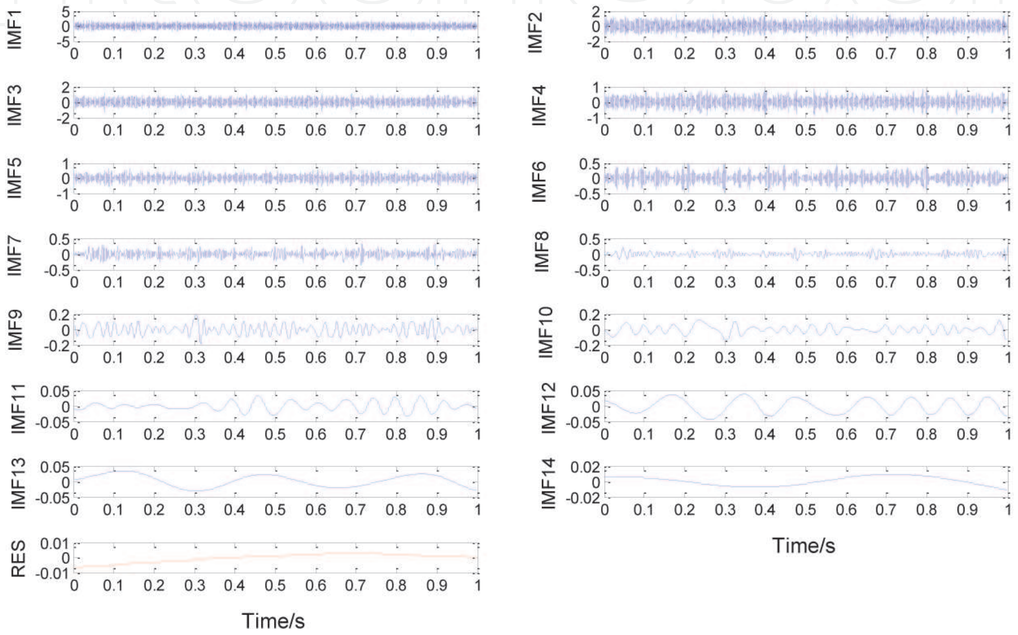

Figure 10. The IMF components and residual of the bearing defect signal from the EMD decomposition.

Component	Correlation coefficient	Component	Correlation coefficient	Component	Correlation coefficient
IMF1	**0.7084**	IMF6	0.1348	IMF11	0.0152
IMF2	0.4767	IMF7	0.0913	IMF12	0.0152
IMF3	0.3220	IMF8	0.0535	IMF13	0.0169
IMF4	0.2353	IMF9	0.0485	IMF14	0.0087
IMF5	0.1747	IMF10	0.0340	RES	−0.0048

Table 3. The correlation coefficients of the IMF components and the residual.

It is shown in the figure that the first IMF component (IMF1) has the highest correlation coefficient with the original signal implying that it is most closely related to the defect signal. Therefore, an envelope analysis is undertaken on the IMF1 component in the next step of analysis. The result is shown in **Figure 11** where the bearing defect frequency and its second-order harmonic can be clearly discriminated from the envelope spectrum.

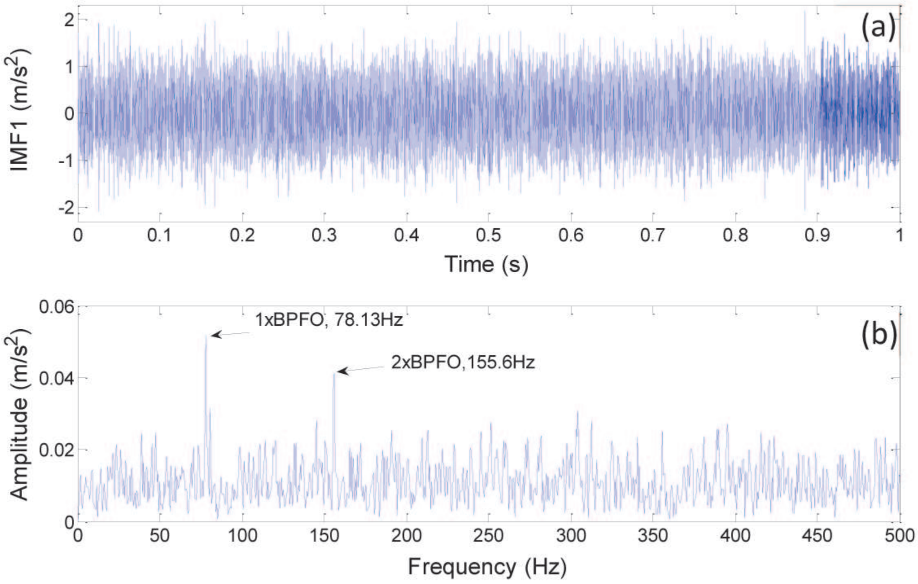

Figure 11. The envelope spectrum of the first IMF component of the bearing defect signal.

Case 4 (Signal de-noising using SVD decomposition): The singular-value decomposition described in Section 3.1.4 can be employed to filter out the noise from bearing condition-monitoring signals to enhance the impulses produced by a bearing defect. In this approach, the one-dimension time waveform of the bearing condition-monitoring signal $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ (as shown in **Figure 12(a)**) is rearranged by an incrementing process to form a Hankel matrix $\mathbf{A}(p, g)$ which is then decomposed into three matrices using Eq. (9) in Section 3.1.4 to obtain the singular values $\sigma_i,\ i = 1, 2, \ldots L,$ whose values are shown in **Figure 13(a)**.

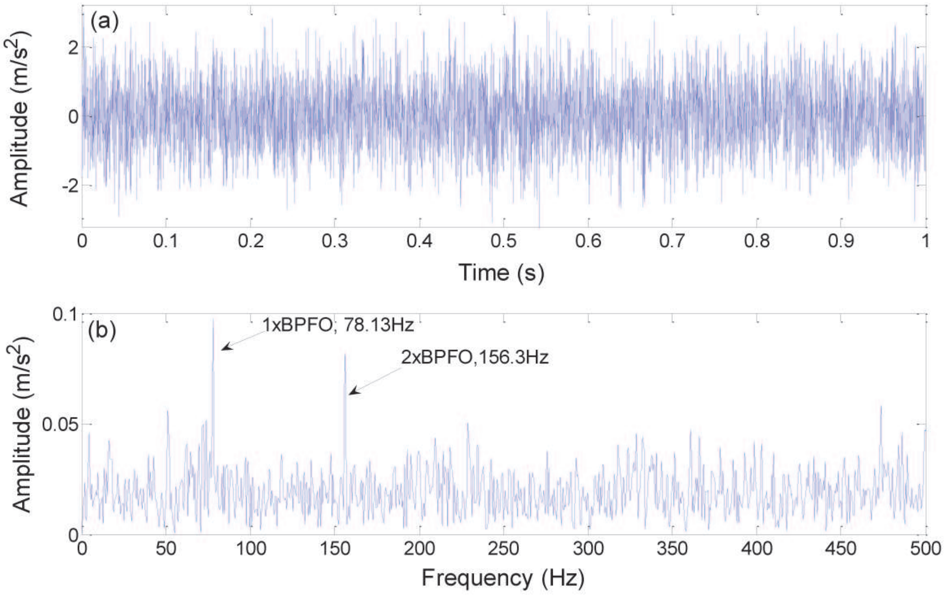

Figure 12. Noise-added bearing defect signal and its envelope spectrum: (a) time waveform (note: the sampling frequency in this simulation is reduced to 5 kHz to avoid the requirement of large computer memory in SVD decomposition; (b) envelope spectrum.

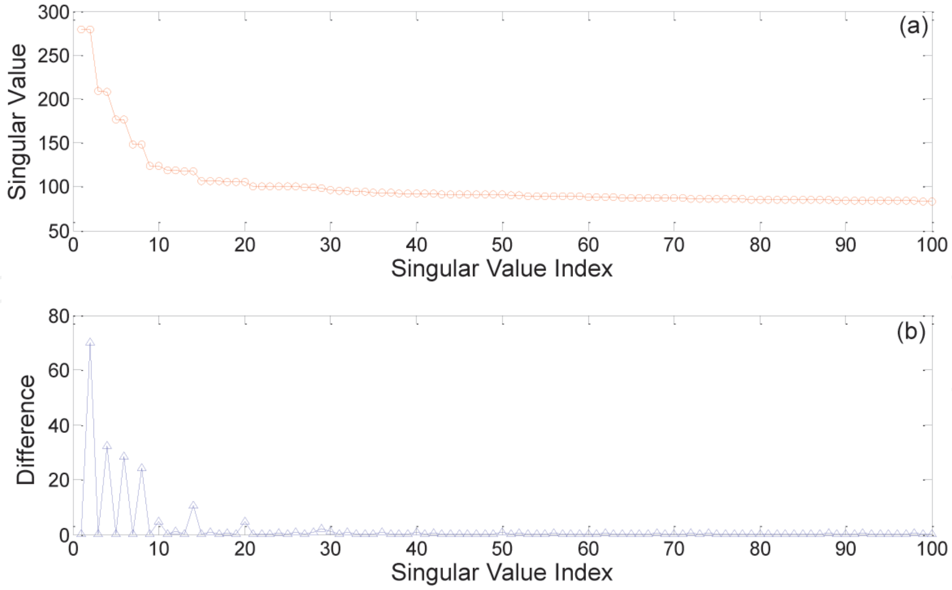

Figure 13. (a) Singular values and (b) the series of the difference between two sequential singular values.

The next step is to calculate the difference between the subsequent singular values $b_i = \sigma_i - \sigma_{i+1},\ (i = 1, 2, \cdots, L-1)$ to form a series vector $\mathbf{B} = (b_1, b_2, \ldots, b_{L-1})$ which reflects the variation of the two neighboring singular values. Vector **B** is plotted in **Figure 13(b)** for illustration. When the difference between two neighboring singular values is large, they will form a larger peak in the difference vector **B** (see **Figure 13(b)**) indicating that there is a small correlation between the defect and noise signals at the corresponding singular values. Keeping the singular values prior to as well as the largest difference peak and letting the singular values after this peak to zero and then substituting the modified singular-value vector to Eq. (9) to obtain a new Hankel matrix (note, as the first difference peak happens to be the largest peak in our case, the singular values associated with the sequential three peaks are also used in the modified singular value vector). We can then reverse the process of the first step to obtain the de-noised bearing defect signal as shown in **Figure 14(a)**. The envelope spectrum of the de-noised signal (**Figure 14(a)**) is shown in **Figure 14(b)**. It is shown that the SVD de-noise can yield a much clean spectrum mainly containing the bearing defect frequency component and its higher harmonics.

3.4. Statistical-based bearing fault diagnosis

3.4.1. Statistical features in the time domain

Some useful statistical features obtained directly from a time waveform signal can be used to evaluate the health condition of a rolling element bearing. Such features can be grouped into two categories: (a) dimensional features and (b) nondimensional features. The group of dimensional statistical features includes peak value, root-mean-square (RMS) value, absolute mean value and variance. This feature group is closely related to bearing fault severity, for instance, their values increase as the bearing fault severity increases. Though, care needs to be taken as their values are also influenced by the working conditions such as load or rotate speed of a machine.

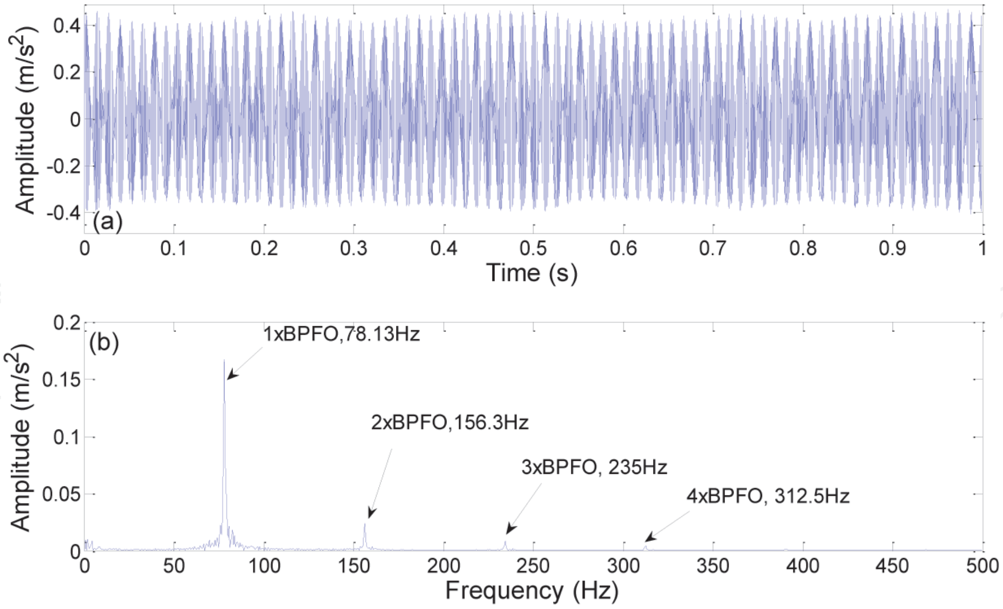

Figure 14. (a) The de-noised bearing defect signal after SVD decomposition; (b) the envelope spectrum.

For a smooth and ergodic continuous time-domain signal $x(t)$, its peak value can be calculated as

$$x_p = \text{Max}[|x(t)|]. \tag{34}$$

Its RMS value which reflects the power level of the signal is calculated by

$$x_{\text{rms}} = \sqrt{\int_{-\infty}^{+\infty} x(t)^2 p(x) dx}, \tag{35}$$

where $p(x)$ is the probability density function of the signal $x(t)$, which represents the probability level that the signal $x(t)$ fall into a certain interval.

Alternatively, an approximate formula can be used to calculate the RMS value as

$$x_{\text{rms}} = \sqrt{\frac{1}{T}\int_0^T x^2(t)dt}. \tag{36}$$

The absolute mean value is defined as

$$x_{\text{av}} = \int_{-\infty}^{+\infty} |x| p(x) dx. \tag{37}$$

Or it can be calculated using the approximate formula:

$$x_{\text{av}} = \frac{1}{T}\int_{-\infty}^{+\infty} |x(t)| dt. \tag{38}$$

Variance is used to depict the fluctuation of a signal that deviated from the center, which can be viewed as the dynamic feature of a signal. The variance of a signal $x(t)$ is

$$D_x = \sigma_x^2 = \int_{-\infty}^{+\infty} (x - \mu_x)^2 p(x) dx, \tag{39}$$

where μ_x is the mean value and σ_x is the standard deviation of the signal.

Variance can also be calculated approximately using the following formula:

$$D_x = \sigma_x^2 = \frac{1}{T} \int_0^T \left(x(t) - \mu_x \right)^2 dt. \tag{40}$$

Table 4 lists the mathematical formula for calculating the same features for a corresponding discrete time waveform, $\{\mathrm{x}(n) | n = 1, 2, \cdots, N\}$.

Statistical features	Formula
RMS	$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i$
Mean	$x_{\text{rms}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2}$
Absolute mean	$x_{\text{av}} = \frac{1}{N} \sum_{i=1}^{N} \lvert x_i \rvert$
Variance	$D_x = \sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - x_{\text{av}})^2$

Table 4. The dimensional statistical features of a discrete time series.

The group of nondimensional statistical features includes crest factor, shape factor, impulsion factor, clearance factor, skewness, and kurtosis factors which are the ratio of two-dimensional statistical features. This type of features is insensitive to the change amplitude or frequency in a signal and thus is not influenced by the working condition of a machine and can accurately reflect a fault condition of a bearing. The formulas for these features are listed in **Table 5**:

Statistical features	Formula
Crest factor	$C = \frac{x_p}{x_{\text{rms}}}$
Shape factor	$S = \frac{x_{\text{rms}}}{x_{\text{av}}}$
Impulsion factor	$I = \frac{x_p}{x_{\text{av}}}$
Clearance factor	$Cl = \frac{x_p}{x_r}$
Skewness	$Sk = \frac{\int_{-\infty}^{+\infty} \left(x(t) - \mu_x \right)^3 p(x) dx}{\sigma_x^3}$(continuous) $Sk = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^3}{(N-1)S^3}$(discrete)
Kurtosis	$K = \frac{\int_{-\infty}^{+\infty} \left(x(t) - \mu_x \right)^4 p(x) dx}{\sigma_x^4}$(continuous) $K = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^4}{(N-1)S^4}$(discrete)

Table 5. The nondimensional features of a time waveform.

In **Table 5**, x_r is the root amplitude of a continuous time waveform, which is given by

$$x_r = \left(\int_{-\infty}^{+\infty} |x(t)|^{1/2} p(x)dx\right)^2. \tag{41}$$

Its discrete form is given by

$$x_r = \left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{|x_i|}\right)^2. \tag{42}$$

Shape factor and impulsion factor are often used to examine whether there exists pulse shocks in a signal. Clearance factor is sometimes used to examine the wear condition of a machine. Yet, the most frequently employed nondimensional features in data analysis are the skewness and kurtosis factors. The skewness is the third statistic moment that measures the degree of asymmetric distribution of a time waveform. The kurtosis is the fourth statistic moment that measures the "Peakness" of the data distribution.

3.4.2. Statistical features in the frequency domain

The power spectrum depicts the power amplitude level of the frequency components in a signal. If the power amplitude levels for some of the frequency components change, the weight-averaged center frequency of a power spectrum will also change. For example, if the number of frequency components with dominant amplitude in a spectrum increases, the energy distribution will be more dispersed. On the contrary, the energy distribution will be concentrated more around the dominant frequency components if there are only a few of such components in the spectrum. Hence, the fluctuation of a signal in the frequency domain can reflect the change of machine condition by observing the change in the weight-averaged center frequency of power spectrum or the disperse degree of power amplitude level distribution.

The typical statistical featured used to depict the weight-averaged center of a power spectrum are "Frequency center" and "Mean-square frequency," which are defined as follows:

$$\text{Frequency center} : \text{FC} = \frac{\int_{-\infty}^{+\infty} f S(f)df}{\int_{0}^{+\infty} S(f)df}, \tag{43}$$

and

$$\text{Mean-square frequency} : \ \text{MSF} = \frac{\int_{0}^{+\infty} f^2 S(f)df}{\int_{0}^{+\infty} S(f)df}, \tag{44}$$

where $S(f)$ represents the power spectrum of a continuous waveform $x(t)$.

The statistical feature used to depict the disperse degree of energy distribution in the frequency domain is the "Variance of frequency," which is defined as

$$\text{Variance of frequency : VF} = \frac{\int_0^{+\infty}(f{-}FC)^2 S(f)df}{\int_0^{+\infty}S(f)df} = \text{MSF}{-}\text{FC}^2. \tag{45}$$

Accordingly, for a discrete time series $\{\mathrm{x}(n)|n = 1, 2, \cdots, N\}$, the three frequency-domain statistical features are given by

$$\text{FC} = \frac{1}{2\pi\Delta}\frac{\int_0^{\pi}\omega S(\omega)d\omega}{\int_0^{\pi}S(\omega)d\omega}, \tag{46}$$

$$\text{MSF} = \frac{1}{4\pi^2\Delta^2}\frac{\int_0^{\pi}\omega^2 S(\omega)d\omega}{\int_0^{\pi}S(\omega)d\omega}, \tag{47}$$

and

$$\text{VF} = \frac{1}{4\pi^2\Delta^2}\frac{\int_0^{\pi}(\omega{-}2\pi\Delta FC)^2 S(\omega)d\omega}{\int_0^{\pi}S(\omega)d\omega} = \text{MSF}{-}\text{FC}^2, \tag{48}$$

where Δ is the sampling frequency, $S(\omega)$ is the power spectrum of a discrete time series $x(n)$, which can be obtained using the following formula:

$$S(\omega) = X(\omega)\overline{X(\omega)}, \tag{49}$$

and

$$X(\omega) = \sum_{n=1}^{N-1} x(n)\mathrm{e}^{-\mathrm{j}\pi\omega}, \tag{50}$$

where ω is the angular frequency.

3.4.3. Data complexity index

The complexity of a signal can be described by two measures: entropy and Lempel-Ziv complexity.

Entropy is a measure of randomness, suggested by Shannon in 1948 [21]. Entropy is used to depict the randomness (or "uncertainty") existed in a signal or the amount of information carried by the signal.

For a discrete random variable x with probability density function $\{x_i|i = 1, 2, \cdots, N\}$, the Shannon entropy is given by

$$H(x_i) = -\sum_{i=1}^{n} p(i)\log_{\mathrm{b}} p(i), \tag{51}$$

where $p(i)$ is the probability of the ith event of the random variable x, b is the base of the logarithm which takes the common logarithm base values of either 2, e, or 10 depending on the application. For events with a probability around 0 or 1, the term $p(i)\log_b p(i)$ converges to zero and the entropy is zero. For random signals with uniform probability density function such as pure noise, the entropy is maximum.

The entropy-based techniques have been widely used in bearing fault diagnosis in the last decade. It has been shown that the entropy value is closely related to the working condition of a machine and its value decreases monotonously with aggravation of faults or conditions [22]. Entropy is often combined with other techniques to capture the detail changes of the nonlinearity and nonstationary properties of a signal in machine fault diagnosis [23, 24]. Typical entropies used in machine fault diagnosis are approximate entropy, sample entropy, fuzzy entropy and permutation entropy. Yet, some of these features are still not ideal and problematic in CM applications. For example, approximate entropy is heavily dependent on the data record length which could yield lower estimation value. Sample entropy uses Heaviside step function which is mutational and discontinue at the boundary. Fuzzy entropy is calculated based on the membership function which is difficult to determine accurately. Permutation entropy requires the reconstruction of phase space though the embedding dimension and time lag of the reconstructed matrix need to be selected manually which has so far limits its application.

Ziv and Lempel [25] presented a specific complexity algorithm termed as Lempel-Ziv complexity (LZC) to calculate the complexity of a finite length time series. LZC can reflect the rate for generating the new condition pattern feature as the nonlinearity of a time series grows [26]. Its value represents the degree of random variation of a time series. In their algorithm, the complexity values of a time series is evaluated based on a ‘‘coarse-graining’’ operation by which the data sequence is transformed into a pattern of only a few symbols, for example, 0 and 1 and involves data sequence comparison and number counting in one dimension only. A flow chart of the LZC algorithm is shown in **Figure 15** and the process is described below:

1. Coarse-grain process to a finite binary sequence. A discrete time series $A = \{a_1, a_2, \cdots, a_n\}$ is converted into a binary sequence $S_N = \{s_1, s_2, \cdots, s_n\}$ in the initiation and preprocessing phase.

2. Copy and Insert. A binary sequence up to $s_r(1 < r < N)$ of complexity c_N can be reconstructed by simply copying and inserting some of the existing vocabulary of $SQv_r = \{s_1, s_2, \cdots, s_v\} \quad (v < r)$. To check the rest string $S_{N-r} = \{s_{r+1}, \cdots, s_N\}$ can be reproduced by the same approach, the process is executed by the following steps:

   Step 1: Take $Q_r = \{s_r\}$ and check whether this string belongs to the vocabulary of SQv_r. If so, string $Q_r = \{s_r\}$ is a simple repetition of an existing substring of SQv_r (i.e., a simple “copy” of the existing vocabulary can restore it) and the complexity remains unchanged or $c_N(r) = c_N(r–1)$.

   Step2: Read the next string and take $Q_{r+1} = \{s_r, s_{r+1}\}$. Check if $Q_{r+1} = \{s_r, s_{r+1}\}$ belongs to SQv_{r+1}.

   Step 3: If string Q_{r+1} does not belongs to SQv_{r+1}, increase the complexity by one, i.e., $c_N(r+1) = c_N(r) + 1$, nullify $Q_{r+1} = \{\}$, read the next string and take $Q_{r+3} = \{s_{r+3}\}$.

Step 4: Repeat the above process until $S_N = \{s_1, s_2, \cdots, s_n\}$ is covered. The resulting $c_N(n)$ is the complexity of a given string.

3. Normalization of the complexity value. The complexity obtained above equals the number of nullification of Q. It indicates that the complexity is affected by the length of the string, or the number of data sample N. To find a robust complexity measure, Lempel and Ziv [27] suggested a normalized measure $c_{LZ} = \frac{c_N(n)}{b_N(n)}$, which is termed as LZC after their names and defined by

$$0 \le c_{LZ} = \frac{c_N(n)}{b_N(n)} \le 1 \tag{52}$$

where $b_N(n) = \lim_{n \to \infty} c_N(n) \approx \frac{n}{\log_k n}$.

It is shown in **Figure 15** that the coarse-grained process of a finite time sequence serves as the basis for the LZC algorithm. Commonly employed technique in coarse-grained process is a single-scale process which converts a discrete time series $A = \{a_1, a_2, \cdots, a_n\}$ into a symbolic binary sequence $S_N = \{s_1, s_2, \cdots, s_n\}$ using the following formula:

$$s_i = \begin{cases} 1, & a_i \ge \overline{a} \\ 0, & a_i < \overline{a} \end{cases} \tag{53}$$

where $\overline{a} = (a_1 + a_2 + \cdots + a_n)/n$ is the mean value of the discrete time series which is used as a threshold in the coarse-graining preprocess.

In the single-scale coarse-grain process, all segments of a discrete time series larger than or equal to the mean value are set to 1 while all segments smaller than the mean value are set to 0. The process neglects the fluctuation between the data intervals and loses many detailed information of the data series during the process. In order to capture the details contained in a discrete time series, the division interval should be reduced and a multidivision scale should be adopted so that the variation in the data can be reflected in the binary sequence at a multiscale process. The preprocess of a revised multiscale coarse grain is outlined below:

1. Divide the discrete time series into various scales. A two-scale division process is used here as an example. In this process, a discrete time series is divided into two regions, with the mean value of each region as the boundary. The same approach can be applied to a multiscale division where a discrete time series can be divided into several regions.

2. When the first number in the discrete time series is larger than the mean value of the entire discrete time series, this point is set to 1 or 0 vice versa. Starting from the second data point of the discrete time series, the binary value is determined by comparing its value with the value of the previous point in the discrete time series. If the two points are in the same division interval, the binary value of the latter point will be the same as the previous one. When the value of the latter point increases to another division interval, the point will

be assigned to the binary value of 1. When the value of the latter point decreases into another division interval, the point will be assigned to the binary value of 0.

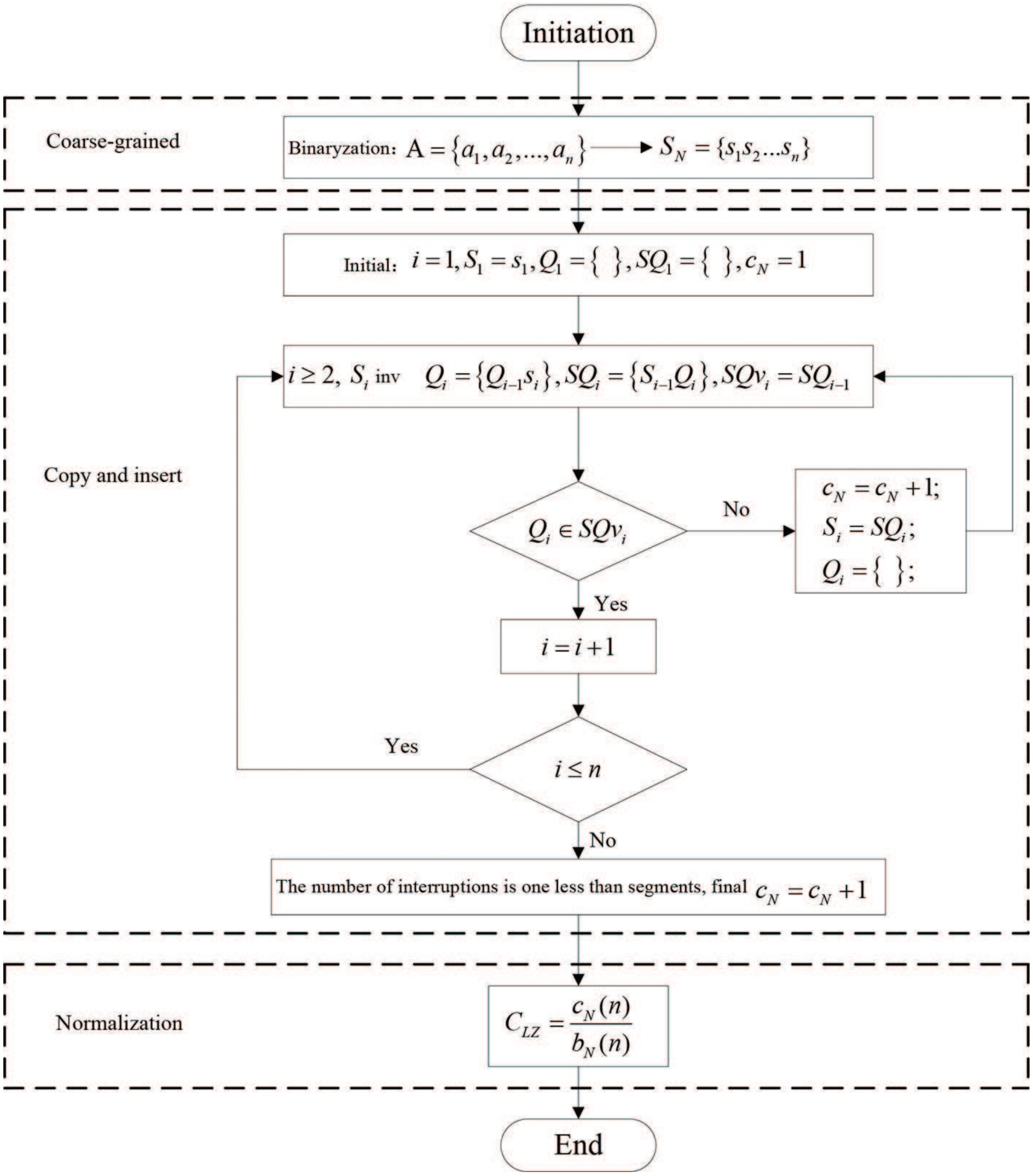
Initiation
Coarse-grained
Binaryzation: $A=\{a_1,a_2,\ldots,a_n\}$ $\rightarrow$ $S_N=\{s_1s_2\ldots s_n\}$
Copy and insert
Initial: $i=1, S_1=s_1, Q_1=\{\ \}, SQ_1=\{\ \}, c_N=1$
$i\geq 2$, S_i inv $Q_i=\{Q_{i-1}s_i\}, SQ_i=\{S_{i-1}Q_i\}, SQv_i=SQ_{i-1}$
$Q_i\in SQv_i$
No
$c_N=c_N+1$; $S_i=SQ_i$; $Q_i=\{\ \}$;
Yes
$i=i+1$
Yes
$i\leq n$
No
The number of interruptions is one less than segments, final $c_N=c_N+1$
Normalization
$C_{LZ}=\frac{c_N(n)}{b_N(n)}$
End

Figure 15. Flow chart of the LZC algorithm.

A data sample given by **Figure 16** is used here as an example to illustrate the difference between the single-scale and multiscale coarse-grain process. For single-scale coarse-grain process, the binary sequence of the data sample is $S_N = (1{,}1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0)$ which does not reflect the fluctuation between the data interval of the time series. When a two-scale coarse-grain process is employed to process the same data, the binary sequence becomes S_N=(1,1,0,1,0,1,0,1,0,1,0,1). Therefore, a multiscale coarse-grain process can better capture the detailed fluctuation in a discrete time series.

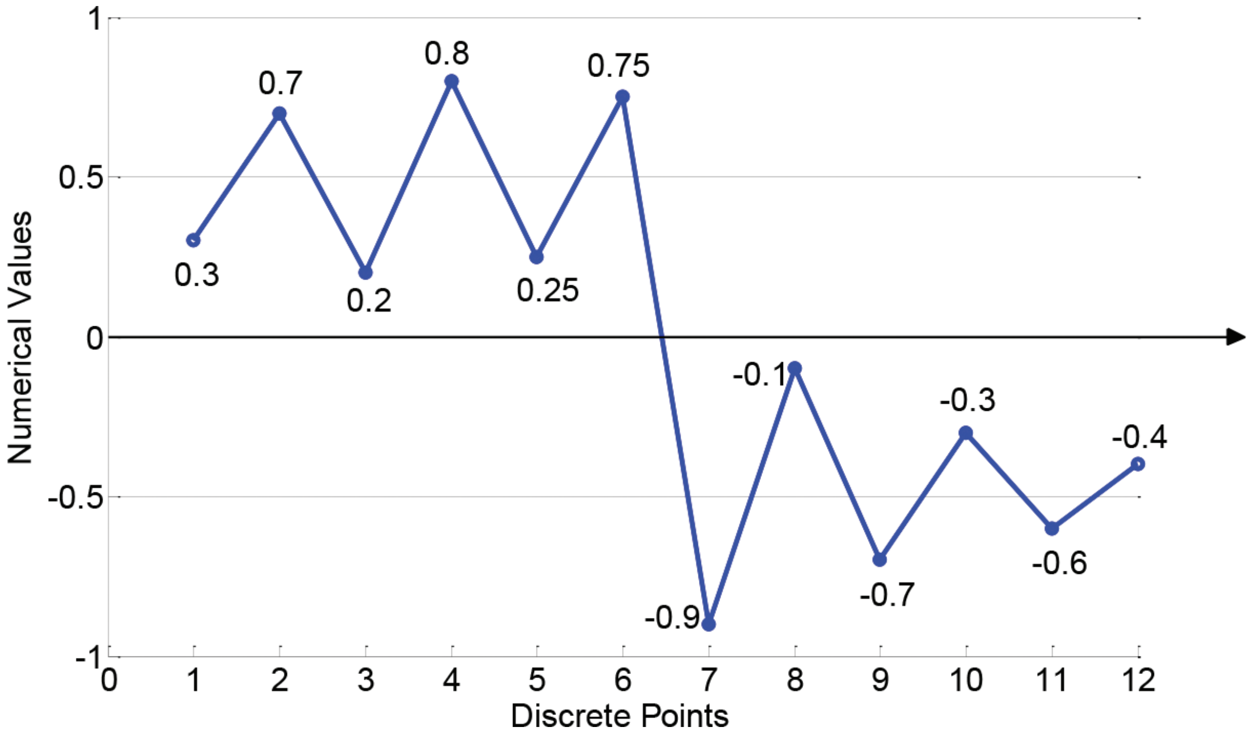

Figure 16. An example of a discrete time series.

3.4.4. Manifold learning

Classical dimensionality reduction techniques, such as multidimensional scaling (MDS) and independent component analysis (ICA), are only applicable to linearly structured data sets but not suitable for high-dimensional, nonlinear data sets such as bearing CM data. To overcome this problem, a nonlinear dimensionality reduction technique, manifold learning, has recently developed for machine fault diagnosis [28]. Manifold learning technique projects the original high-dimensional data onto a lower-dimension feature space while preserving the local neighborhood structure to detect the intrinsic structure of nonlinear high-dimensional data. Manifold learning can be realized through several algorithms including locally linear embedding (LLE), isometric feature mapping (IsoMap), local tangent space alignment (LTSA) and local preserving projection (LPP).

The application of manifold learning in mechanical fault diagnosis can be in twofolds. Firstly, fault features with a large dimension can have many redundant components, which can increase the complexity and operation time of a fault diagnosis process. Manifold learning can be used to eliminate the redundant components and extract the nonlinear features for fault classification in this case. Secondly, manifold learning can discard the noise components and extract the intrinsic manifold features related to nonlinear dynamic of a CM signal; therefore, it can also be used as a de-noise technique. It should be noted that manifold learning only operates on a matrix, so a preprocessing such as a reconstruction of phase space converting an original one-dimensional signal into a two-dimensional data is required. For a detailed information of the manifold learning algorithm and its application in fault diagnosis, interested readers are referred to [28, 29].

3.5. Bearing fault diagnosis based on artificial intelligent

3.5.1. Shallow architecture machine learning

The most commonly employed machine-learning techniques in fault diagnosis of rotating machines are hidden Markov models (HMMs), support vector machines (SVMs) and artificial neural networks (ANNs), which exploit shallow architectures either contain a single hidden layer or without a hidden layer. Such shallow architecture-based models have achieved great success both in theory and in practical applications. Though, shortcomings of these algorithms such as poor universality, lacking of theory basis in parameter selection, easier to fall into a local optimum value have limited the application of the algorithms in machine fault diagnosis.

3.5.2. Deep neural network

A consensus criterion for an effective bearing fault diagnosis technique is that it should not only be able to identify various bearing fault conditions but also be able to discriminate different fault severities in each fault condition [30]. This leads to a stricter requirement on identification procedures where a classifier must have a greater capability in discriminating different fault classes. When fault data contain more than one level of fault severities in each fault condition, the accuracy of fault diagnostic result using shallow architecture classifiers described in the previous section will reduce dramatically. Hinton and Salakhutdinov [31] proposed a deep learning technique to overcome the deficiencies of single-layer architecture classifiers for a better pattern recognition capability. The concept is further extended to become a so-called deep belief networks (DBNs) [32] which relieves the training difficulties of deep network structures by adopting a layer-by-layer unsupervised forward pretraining learning and then back fine-tuning mechanism. A description of the training process of a DBN is given in **Figure 17**.

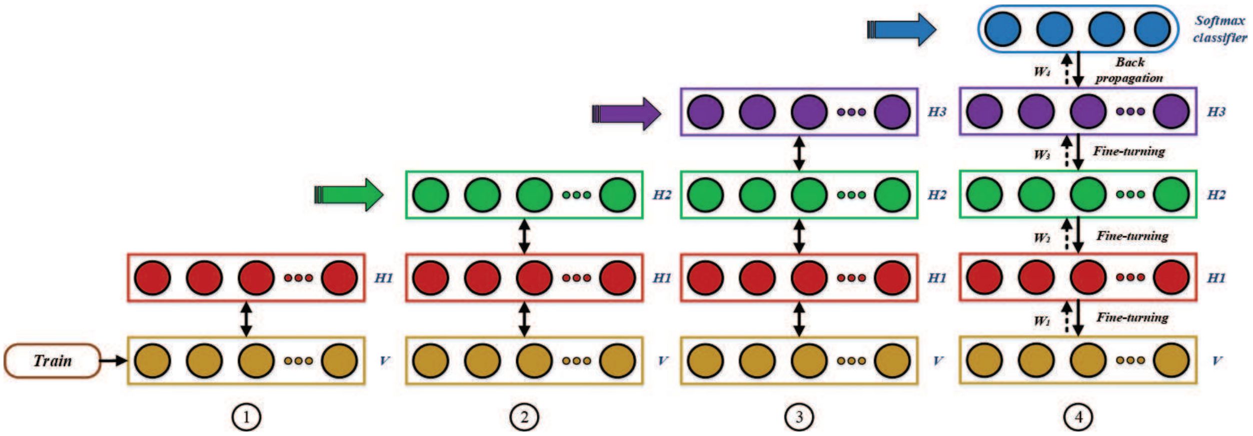

Figure 17. A description of the training process of a DBN.

Bengio et al. [33] proposed a deep stack auto-encoder (SAE) network by using a network structure similar to DBNs which is stacked with a number of auto-encoder networks. Furthermore, Le Cun et al. [34] proposed a convolutional neural network (CNN), a multilayer network where each layer is composed of several two-dimensional planes, to reduce the number of parameters in the learning process using a unique weight-sharing mechanism. The above cited deep-learning-based neural network algorithms have been widely employed in machine fault diagnosis nowadays [35, 36].

3.5.3. A case study on bearing fault diagnosis

Case 5: In this case study, a combination of a four-level wavelet packet decomposition (WVD), a locality preserving projection (LPP), a particle swarm optimization (PSO) and support vector machine (SVM) algorithms are employed for an intelligent bearing fault diagnosis and recognition. Bearing condition-monitoring data on various operation conditions such as healthy bearing, outer race fault, inner race fault and ball fault acquired from a bearing fault simulation test-rig as shown in **Figure 18** are used in this analysis.

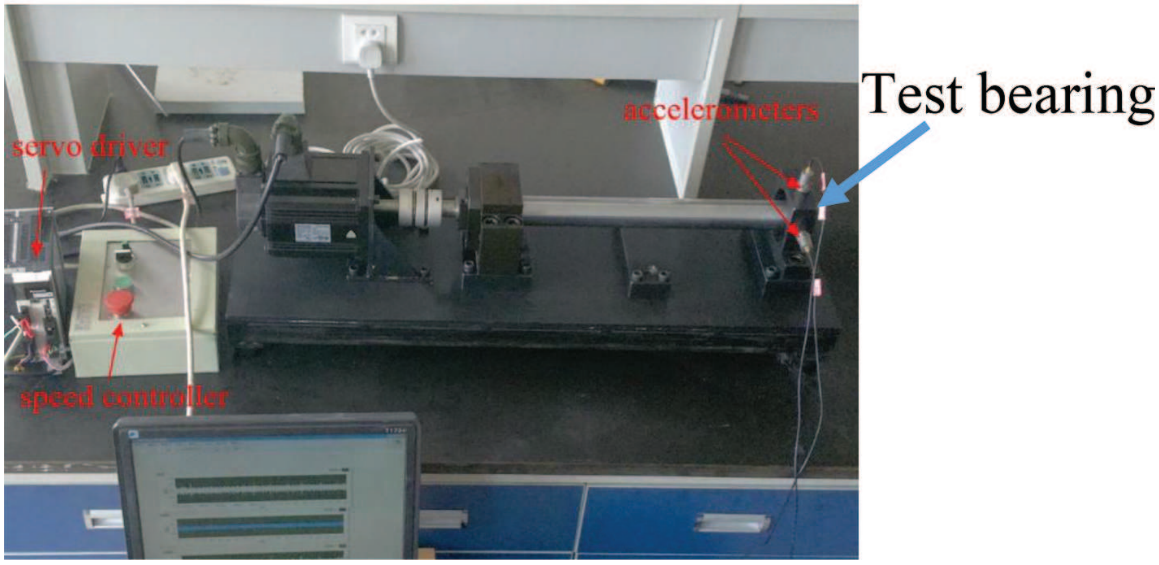

Figure 18. A graphical illustration of the bearing fault simulation test rig.

Three types of bearing defects are simulated in this experiment: an outer race fault, an inner race fault and a ball fault as shown in **Figure 19**. The measured vibration signals from an accelerometer mounted on the test bearing house for the four bearing operation conditions (the three simulated faults and a healthy bearing) are shown in **Figure 20**. Hundred data sets are acquired in the experiment which are divided into two groups: one contains 70 sets of data and is used as the training data set and the other contains 30 sets of data which is used as the test data set.

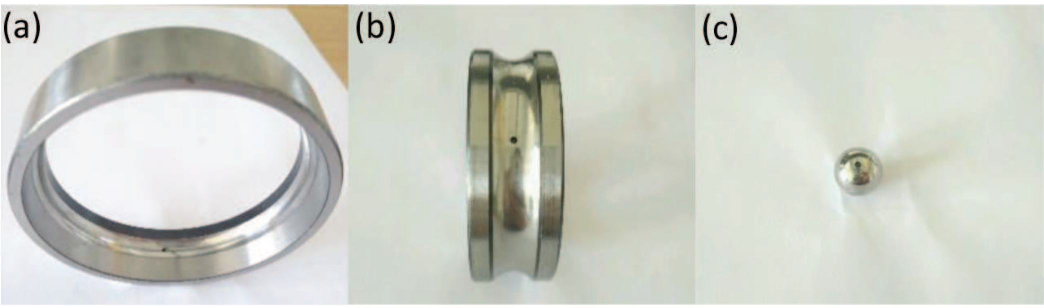

Figure 19. Simulated bearing faults: (a) outer race fault; (b) inner race fault; and (c) ball fault.

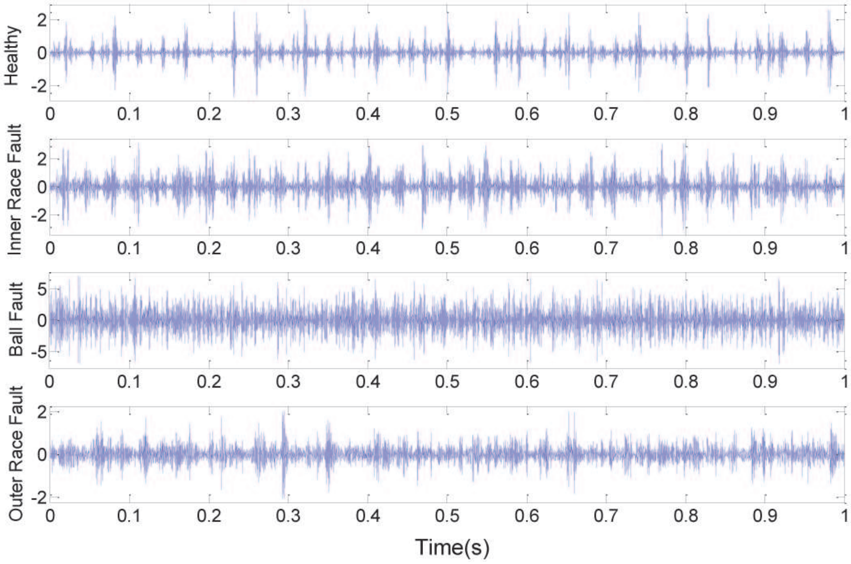

Figure 20. The measured vibration signals for the four bearing operation conditions. Note, sampling frequency, 8192, data length, 8192.

Five major steps are taken in the analysis of the bearing condition-monitoring data:

1. Each data is decomposed by a four-level wavelet packet decomposition leading to 16 components of different frequency contents (mutual orthogonal subspaces). The wavelet components of the condition-monitoring data for the outer race fault are shown in **Figure 21** for illustration.

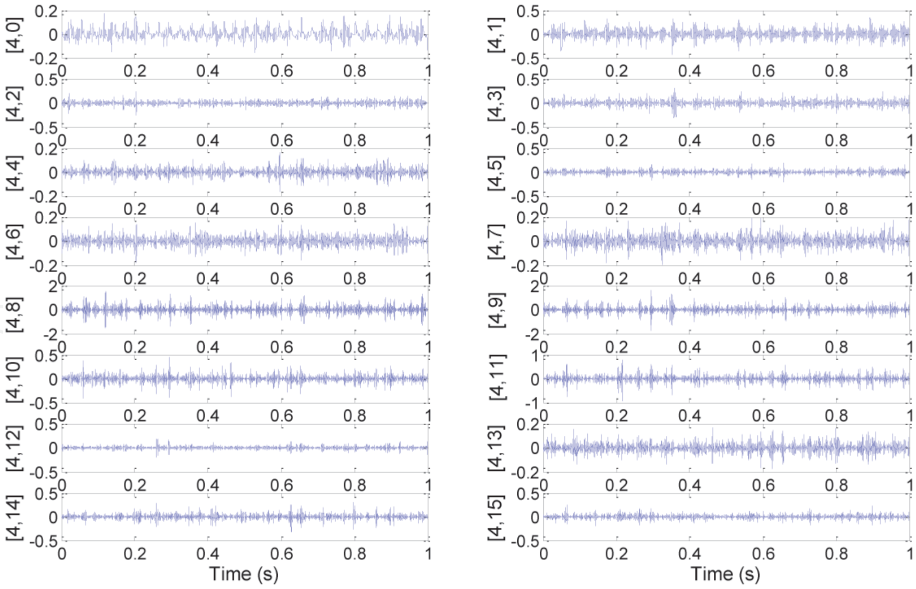

Figure 21. The wavelet components of the outer race bearing defect signal.

2. The energy value for each wavelet component is calculated and is used as the fault feature since the energy of the component can discriminate different classes and contains the fault information and its fluctuation in the particular component corresponding to the occurrence of the fault. A fault feature vector containing 16 energy features can be obtained for

each data. The energy distribution of the 16 wavelet components for the four bearing operation conditions is shown in **Figure 22**.

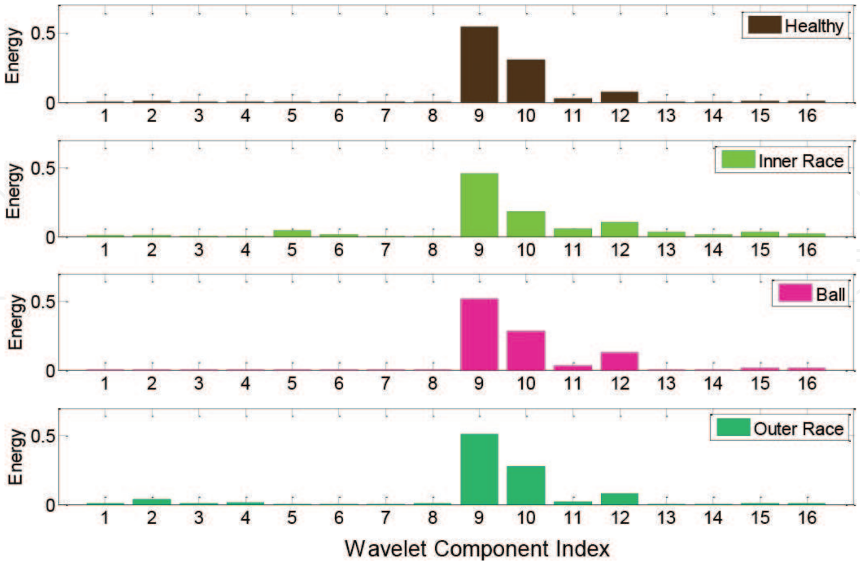

Figure 22. The energy values of the 16 wavelet components for the four bearing operation conditions.

3. For the 100 data sets (of four bearing operation conditions), the process in Step (2) will generate a 400×16 feature set. The dimension of the large feature set can cause a problem for the following algorithm leading to misclassification of the bearing fault classes in the diagnosis step using SVM algorithm. The locality preserving projection (LPP) [37], a linear dimensionality reduction algorithm, which can effectively reduce the dimension while preserving the neighborhood structure of a data set, is utilized to reduce the feature set dimension to 400×3. The three-dimensional feature distribution for the four bearing operation conditions after the dimensional reduction by LPP is shown in **Figure 23**.

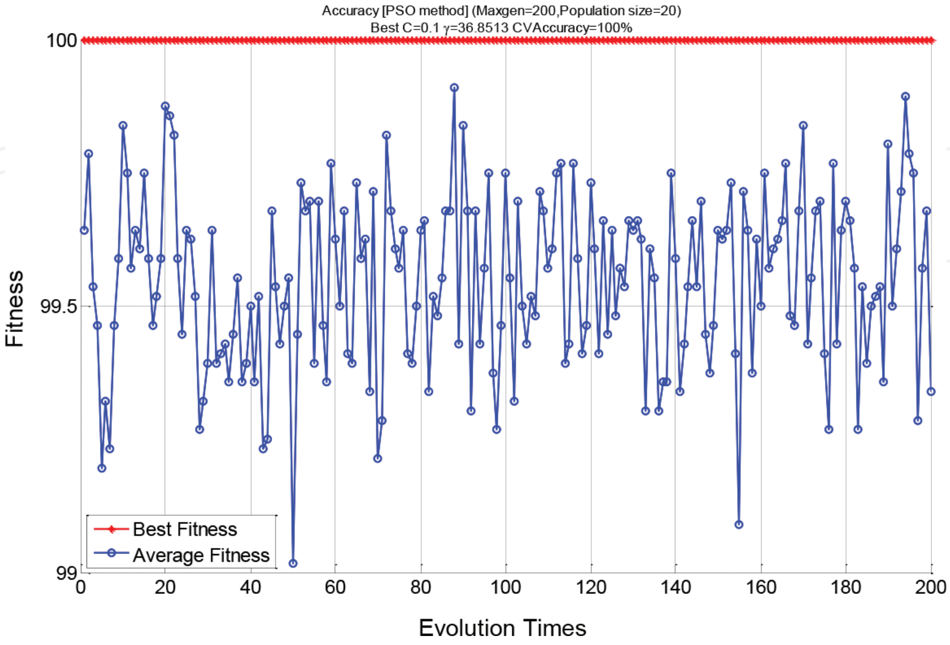

Figure 23. The optimization result of the PSO algorithm.

4. Particle swarm optimization (PSO) is adopted in this step to find the optimal kernel parameters and penalty factor used in SVM algorithm from the training data set to enhance the performance of the SVM algorithm. The optimization result of the PSO is shown in **Figure 23** where the fitness function of the optimization is above 99% (close to zero misclassification rate) throughout the time evolution progress.

5. Training and prediction (recognition) of the bearing experimental data using the SVM algorithm. The prediction result by the SVM is shown in **Figure 24**.

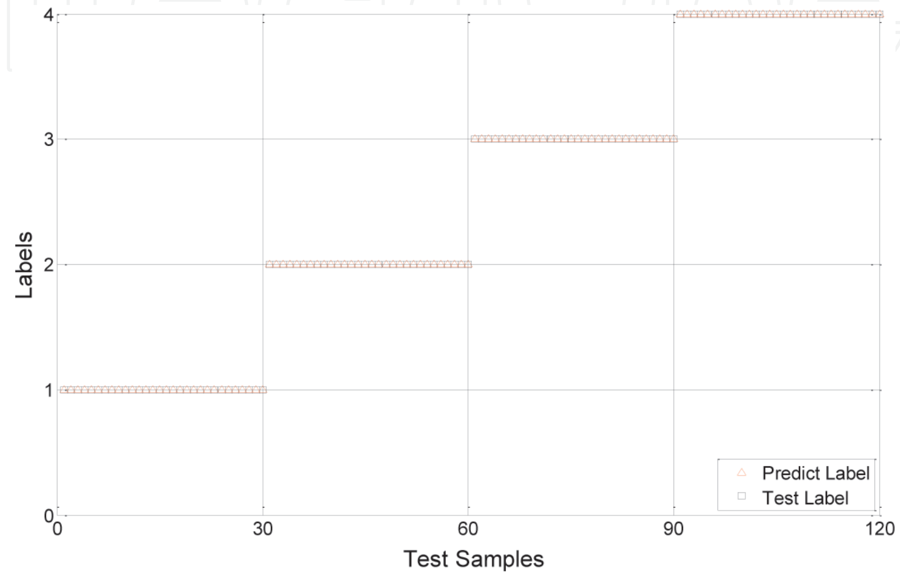

Figure 24. Prediction result of the SVM algorithm. Note, Levels 1–4 correspond to healthy bearing, inner race fault, ball fault and outer race fault, respectively.

4. Conclusion

This chapter presented a comprehensive, step-by-step approach on condition monitoring of roller element bearings aiming to provide an introductory and referencing material for engineers and researchers new to the field. It summarized the most frequently employed data acquisition, signal analysis, feature and parameter extraction and fault diagnosis techniques in the current practice. Pros and cons of each technique are briefly discussed in the text. The formulation and discussion are also supported by ample tables, graphs and figures throughout the text for a better illustration and understanding of how to utilize the techniques presented in the chapter for real-life problems.

Author details

Tian Ran Lin*, Kun Yu and Jiwen Tan

*Address all correspondence to: trlin888@163.com

School of Mechanical Engineering, Qingdao University of Technology, Huangdao District, Qingdao, PR China

References

[1] T. R. Lin, E. Kim and A. C. C. Tan, "A simple signal processing approach for condition monitoring of low speed machinery using Peak-Hold-Down-Sample algorithm", Mechanical Systems and Signal Processing 36(2), 256–270, 2013.

[2] T. R. Lin, W. Wu and A. C. C. Tan, "A signal processing approach to solve the non-linearity problem of acoustic emission sensors", Proceedings of the 9th World Congress on Engineering Asset Management, 28–31 Oct 2014, Pretoria, South Africa.

[3] W. Wu, T. R. Lin, A. C. C. Tan, "Normalization and source separation of acoustic emission signals for condition monitoring and fault detection of a multi-cylinder diesel engine", Mechanical Systems and Signal Processing 64–65, 479–497, 2015.

[4] F. Kadri, F. Harrou, S. Chaabane, Y. Sun, C. Tahon, "Seasonal ARMA-based SPC charts for anomaly detection: Application to emergency department systems", Neurocomputing 173, 2102–2114, 2016.

[5] R. A. Wiggins, "Minimum entropy deconvolution", Geoexploration, 16, 21–35, 1978.

[6] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics—A tutorial", Mechanical Systems and Signal Processing 25, 485–520, 2011.

[7] J. Y. Lee and A. K. Nandi, "Extraction of impacting signals using blind deconvolution", Journal of Sound and Vibration **232** (5), 945–962, 1999.

[8] J. Antoni, "The spectral kurtosis: a useful tool for characterizing non-stationary signals", Mechanical Systems and Signal Processing 20 (2), 282–307, 2006.

[9] J. Antoni, R. B. Randall, "The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines", Mechanical Systems and Signal Processing 20 (2), 308–331, 2006.

[10] Y. Wang, J. Xiang, R. Markert, M. Liang, "Spectral kurtosis for fault detection, diagnosis and prognostics of rotating machines: A review with applications", Mechanical Systems and Signal Processing 66–67, 679–698, 2016.

[11] Y. G. Lei, J. Lin, Z. J. He and Y. Zi, "Application of an improved kurtogram method for fault diagnosis of rolling element bearings", Mechanical Systems and Signal Processing 25, 1738–1749, 2011.

[12] X. Zhao, B. Ye, "Similarity of signal processing effect between Hankel matrix-based SVD and wavelet transform and its mechanism analysis", Mechanical Systems and Signal Processing 23, 1062–1075, 2009.

[13] R. B. Randall, "Vibration-based condition monitoring", John Wiley & Sons, West Sussex, 2011.

[14] K. C. Chua, V. Chandran, U. R. Acharya, C. M. Lim, "Application of higher order statistics/ spectra in biomedical signals—A review", Medical Engineering & Physics 32, 679–689, 2010.

[15] Z. K. Peng and F. L Chu, "Application of the wavelet transform in machine condition monitoring and fault diagnosis: A review with bibliography", Mechanical Systems and Signal Processing 18, 199–221, 2004.

[16] Y. Yin, X. Hu and T. R. Lin, "A practical approach to analyze the non-stationary signals of a quayside container crane motor using a combined empirical mode decomposition and wavelet packet quantization", Noise Control Engineering Journal 64(2), 126–133, 2016.

[17] N. E. Huang, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", Proceedings of the Royal Society, Series A: Mathematical, Physical and Engineering Sciences 454, 903–995, 1998.

[18] Y. Lei, J. Lin, Z. J. He, M. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery", Mechanical Systems and Signal Processing 35, 108–126, 2013.

[19] Z. H. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method", Royal Society of London Proceedings: Mathematical, Physical and Engineering Sciences 460(2046), 1597–1611, 2004.

[20] Z. H. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method", Advances in Adaptive Data Analysis 1(1), 1–41, 2011.

[21] C. E. Shannon, "A mathematical theory of communication", Bell System Technical Journal 27, 379–423, 1948.

[22] H. Cui, L. Zhang, R. Kang and X. Lan, "Research on fault diagnosis for reciprocating compressor valve using information entropy and SVM method", Journal of Loss Prevention in the Process Industries 22, 864–867, 2009.

[23] S. Pan, T. Han, A. C. C. Tan and T. R. Lin, "Fault diagnosis system of induction motors based on multiscale entropy and support vector machine with mutual information algorithm", Shock and Vibration 2016, Article ID 5836717, 2016.

[24] Y. Tian, Z. Wang, C. Lu, "Self-adaptive bearing fault diagnosis based on permutation entropy and manifold-based dynamic time warping" (In Press), Mechanical System and Signal Processing.

[25] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding", IEEE Transactions on Information Theory 24(5), 530–536, 1978.

[26] H. Hong and M. Liang, "Fault severity assessment for rolling element bearings using the Lempel-Ziv complexity and continuous wavelet transform", Journal of Sound and Vibration 320, 425–468, 2009.

[27] A. Lempel and J. Ziv, "On the complexity of finite sequences", IEEE Transactions on Information Theory 22(1), 75–81, 1976.

[28] B. Tang, T. Song, F. Li, L. Deng, "Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine", Renewable Energy 62, 1–9, 2014.

[29] H. S. Seung and D. D. Lee, "The manifold ways of perception", Science 5500(290), 2268–2269, 2000.

[30] M. Gan, C. Wang and C. Zhu, "Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings", Mechanical Systems and Signal Processing 72–73, 92–104, 2016.

[31] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", Science 313(5786), 504–507, 2006.

[32] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing", IEEE Signal Processing Magazine 28(1), 145–154, 2011.

[33] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, "Greedy layer-wise training of deep networks", Advances in Neural Information Processing Systems 19, 153–160, 2007.

[34] Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, "Handwritten digit recognition with a back-propagation network", Advances in Neural Information Processing Systems 2, 396–404, 1990.

[35] Z. Chen, C. Li and R. Sanchez, "Gearbox fault identification and classification with convolutional neural networks", Shock and Vibration 2015, Article ID 390134, 2015.

[36] W. Sun, S. Shao, R. Zhao, R. Yan, X. Zhang and X. Chen, "A sparse auto-encoder-based deep neural network approach for induction motor faults classification", Measurement 89, 171–178, 2016.

[37] X. He and P. Niyogi, "Locality preserving projections", Advances in Neural Information Processing Systems 16 (NIPS 2003), Vancouver, Canada, 2003.