A quantitative estimation method of ball bearing localized defect size based on vibration instantaneous energy analysis

To cite this article: Zhiyuan He *et al* 2022 *Meas. Sci. Technol.* **33** 075011

IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. **33** (2022) 075011 (16pp) https://doi.org/10.1088/1361-6501/ac65da

A quantitative estimation method of ball bearing localized defect size based on vibration instantaneous energy analysis

Zhiyuan He[1,2,3,*], Guo Chen[4,*], Kaiyong Zhang[5] and Konstantinos Gryllias[2,3]

[1] College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, People's Republic of China
[2] Department of Mechanical Engineering, KU Leuven, Celestijnenlaan 300, Heverlee, 3001, Belgium
[3] Dynamics of Mechanical and Mechatronic Systems, Flanders Make, Leuven, Belgium
[4] College of General Aviation and Flight, Nanjing University of Aeronautics and Astronautics, Liyang, 213300, People's Republic of China
[5] Engineering Training Center, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, People's Republic of China

E-mail: cgzyx@263.net and hzy2017@nuaa.edu.cn

Received 25 September 2021, revised 7 December 2021
Accepted for publication 8 April 2022
Published 22 April 2022

Abstract

The quantitative evaluation of bearing damage size is of great significance for accurately predicting the remaining useful life of bearing. Existing studies have shown that rolling elements will produce corresponding vibration features when they enter and leave localized defects, and these features are the premise of accurately estimating the defect size. This paper presents an accurate quantitative estimation method of bearing localized defect size based on vibration instantaneous energy analysis. The time interval of rolling element entry and exit the damaged area is determined by analysing the instantaneous energy of fault impulse. The effectiveness of the proposed method is verified by the experiments of different defect sizes of the outer race and inner race at various speeds. Besides, the proposed method has higher accuracy compared with the conventional methods, which provides a new idea for fast and convenient estimation of the ball bearing localized defect size.

Keywords: rolling bearing, quantitative estimation, defect size, instantaneous energy, smooth pseudo Wigner–Ville distribution

(Some figures may appear in colour only in the online journal)

1. Introduction

Rolling bearing is one of the most common and core parts of rotating machinery, which is widely used in aerospace, railway, construction, steel, paper, textile, mining, and other industries [1]. Its operation state is related to the safety of the whole machine. With recent advances in technology, numerous works of rolling bearings are becoming mature on warning of bearing early failure. For example, classical envelope analysis method [2], various wavelet analysis methods [3, 4], blind deconvolution methods [5–7], signal decomposition methods [8–10], machine learning methods [11, 12] and so on. However, the complete condition monitoring and health management system of the rolling bearing should also include predicting the damage extension and the remaining useful life (RUL) of the bearing. In recent years, although some researches on bearing RUL prediction have appeared one after

* Authors to whom any correspondence should be addressed.

another [13, 14], most of them are evaluated by trend degradation indicators. The key to predicting the RUL of bearings is to quantitatively evaluate the localized damage of bearings accurately. Therefore, this paper will focus on a simple and accurate quantitative evaluation method of bearing localized damage.

Since the 1990s, Epps [15] first proposed to quantitatively evaluate the localized defect of bearings by vibration characteristics, that is, in the process of rolling elements passing through the localized damage, they will produce low-frequency characteristics (entry the defect) and high-frequency characteristics (impact) in the vibration signal. The time interval between these two vibration characteristics is roughly related to the damage size. Based on this theory, many studies began to focus on this issue. They can be divided more or less into two categories. One is based on a combination of multiple signal processing methods. For instance, Sawalhi and Randall [16] proposed two signal processing strategies for localized defect size estimation. One is a joint processing method based on wavelet analysis to enhance entry event and impact event, and the other is a cepstrum-based method to handle two events separately. Afterward, Sawalhi *et al* [17] and Wang *et al* [18] had put forward synchronous averaging and cepstrum editing methods successively. Some signal decomposition methods are also used in this field, such as the method of combining empirical mode decomposition and approximate entropy index proposed by Zhao *et al* [19] extract the entry and exit event independently and Chen and Kurfess [20] combined variational modal decomposition and differentiation techniques to form a method for bearing defects calculation, whose outcomes state clearly that the proposed method has accurate estimation. In recent years, Howard's team has continued to contribute to this field. Petersen *et al* [21] illustrated that with the change of defect size, the stiffness and natural frequency of rigid rolling bearing components will change. The variation of stiffness causes the parametric excitation at the damage frequency, which also produces a low-frequency response at the exit phase. Based on this, corresponding estimation methods of bearing localized defects are proposed by Moazen-Ahmadi *et al* [22, 23] and Larizza *et al* [24]. Recent pieces of literature [25, 26] have begun to pay attention to the signal monitoring of the bearing natural spalling size and compared the effects of vibration monitoring, instantaneous angular speed, acoustic emission, and load monitoring methods, which provides invaluable guiding information for future research. On the other hand, some studies estimated the bearing localized defect size based on the physical-mechanical models. For example, a physical model independent of the load and stiffness parameters of the bearing was proposed in [27], which was solved by numerical integration to calculate the bearing outer race's localized damage size. Kogan *et al* [28] proposed the spalling width expression based on the analysis of the interaction between the rolling element and spalling and then achieved good results [29] in the experimental verification of localized defects on the outer race. At the same time, some methods based on Hertz contact model and the combination of mechanics and signal processing methods have also been reported in [30, 31].

Although these above methods can accurately estimate the bearing damage size under certain conditions, some signal processing methods need to be combined with multiple methods. Too many combinations may make users, especially non-signal processing professional users, feel complicated and inconvenient for theirs engineering application software where the calculation cost also will be very high. Besides, the proposed physical models need to know the inherent parameters of the bearing in advance such as stiffness, Poisson's ratio, modulus of elasticity, and need to calculate maximum load, contact stress, and so on. In addition, the physical model methods need complex numerical integral equations. Also, some estimation methods based on the physical model lack discussion of the inner race localized damage size evaluation. Therefore, it is necessary to put forward a convenient method to evaluate the localized damage size of the inner and outer race so as to be better applied to realistic engineering problems.

In the process of quantitative calculation of bearing localized damage, getting the accurate time nodes of entering and leaving is the premise of exact estimation. In the fault impulses, these characteristics and the vibration frequencies go hand in hand. Therefore, the time-frequency distribution (TFD) methods should be the effective way to solve this problem. For example, short-time Fourier transform [32], continuous wavelet transforms (CWT) [3, 33], Wigner–Ville distribution (WVD) [34], and other methods have been widespread applied in the fault diagnosis' field. But it is worth noting that when analysing the fault impulses caused by the localized defect of bearings, the TFD method with high resolution should be selected. Because the time interval of the ball passing through the damaged area is very short, it is necessary to extract the instantaneous time nodes of entering and leaving exactly to ensure the calculation accuracy of damage size. Based on this, this paper presents a new quantitative estimation method for ball bearing localized defect based on instantaneous energy, which provides a new idea for rapid, simple, and accurate estimation of bearing localized defect size. The time nodes of entry and exit phases can be exactly captured by analysing the variation of instantaneous energy extracted from smooth pseudo WVD (SPWVD). The major works of this study are:

- By analyzing the instantaneous energy of the fault impulse, the time nodes of the entering and leaving phases of the ball passing through the damaged area are accurately extracted. Especially when the low-frequency response is easily disturbed by the high-frequency response in the exit phase, the proposed method can separate them accurately.
- The effects of classical linear TFD (CWT) and quadratic non-linear TFD (SPWVD) in quantitative defect estimation of bearings are compared. The results indicate that SPWVD has more advantages in extracting the details of the low-frequency response. At the same time, the influence of different center frequencies of complex Morlet wavelet on

the results is analyzed; the selection of the optimal window function of SPWVD in quantitative estimation of bearing damage is also discussed.

In this paper, section 2 describes typical localized defect vibration characteristics. Section 3 reviews the SPWVD and the advantages of SPWVD over WVD are explained. Section 4 introduces the specific steps of the proposed method. Experimental verification and results comparisons are shown in section 5. The selection of wavelet parameters and the comparison of SPWVD window functions are discussed in section 6. Section 7 are the conclusions.

2. Description of localized defect vibration characteristics of bearings

Early evidence [15, 16] indicated that there will be a low-frequency response (entering) and high-frequency response (impact) when a ball passes through the defect. Based on this theory, some methods [17–20, 27, 30] are proposed to compute the size of bearing localized damage. However, [21–23, 28, 29] shows that the low-frequency response also occurs when the roller leaves the damage, as the low-frequency feature is not easily detected in the time domain due to their mixing with the high-frequency feature. It can be explained as three phases. At the entry, with the contacts between the ball and the raceways are gradually separated, the contact stress will be reduced to zero, resulting in the de-stressing at the entry-transient phase. As the number of rolling elements decreases, the system mode will be excited in low frequency, resulting in a low-frequency vibration response. The next phase is the moment during the ball impacts the defect trailing edge. The collide will excite huge vibration energy, causing a high-frequency response. Finally, the ball and the raceways re-contact when it exits the defect. At this transient phase, the bearing load is restored, and low-frequency attenuation will occur at this re-stressing moment.

To visualize this process, figure 1 draws the process of a ball entering and leaving the localized defect. Figure 2 plots the vibration response of a ball traversing over the localized damage with a width of 1.4 mm on the outer race. **A, B** and **C** are the phases of entry, impact, and exit, respectively. From figure 1, the defect size is $l = d_1 + d_2$. Therefore, the key to accurately calculate the localized defect size is to get the time interval ΔT between **A** and **C**. However, some previous studies [16–20, 27, 30, 31] have calculated the time interval between **A** and **B**, because the low-frequency (LF) response and impact high-frequency (HF) response are obvious in the time domain. In fact, the result is only the dimension of d_1, so it is not accurate to a certain degree. In the time domain, it is hard to make a distinction to recognize the dividing time node between the high-frequency vibration response of **B** and the low-frequency oscillation of **C**. Therefore, this paper will extract these instantaneous moments from the perspective of the time-frequency domain.

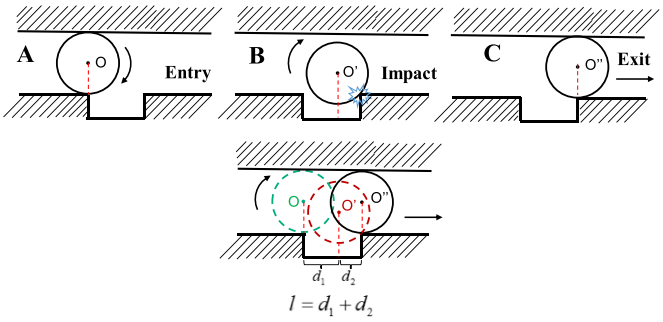


Figure 1. A ball entering and leaving the localized defect.

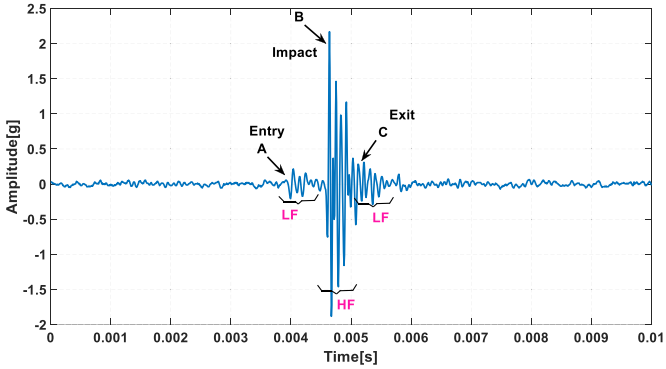


Figure 2. Vibration response of a ball traversing over the localized defect with a width of 1.4 mm on the outer race.

3. Review of SPWVD

The idea of TFD is to establish a function that can describe the energy density distribution of signal using time and frequency at the same time, and given a method to estimate the signal energy density distribution. Thus, a joint time-frequency energy density distribution can be defined as $P(t,\omega)$ which represents the energy density of the signal at time t and frequency ω. Then the total energy of the signal E_{total} is:

$$E_{\text{total}} = \int_{-\infty}^{+\infty}\int P(t,\omega)\,dtd\omega \tag{1}$$

$P(t,\omega)$ must satisfy the following marginal condition,

$$|s(t)|^2 = \int_{-\infty}^{+\infty} P(t,\omega)d\omega \tag{2}$$

$$|S(\omega)|^2 = \int_{-\infty}^{+\infty} P(t,\omega)dt \tag{3}$$

where $|s(t)|^2$ is the instantaneous energy, $|S(\omega)|^2$ is the energy spectrum.

WVD is a bilinear time-frequency analysis method, which is defined as the Fourier transform of signal instantaneous correlation function. Because there is no window function, WVD has high resolution in the analysis of instantaneous frequency

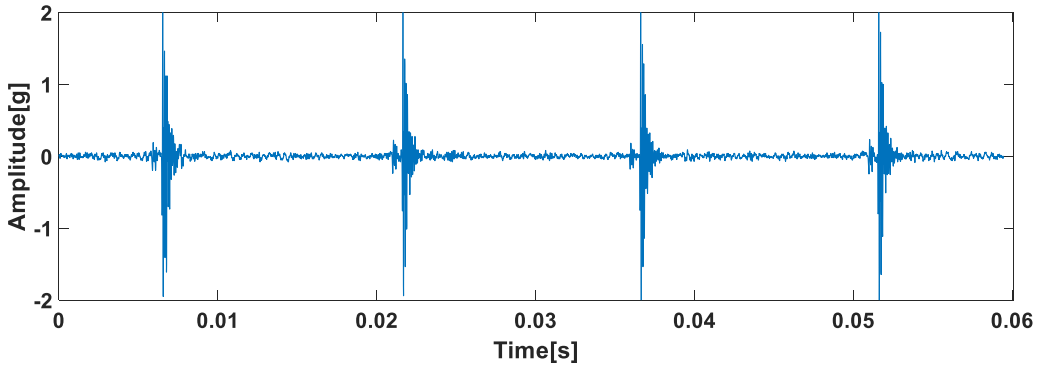


Figure 3. Vibration response of outer race with localized defect of 1.4 mm.

changing with time. The WVD of continuous nonstationary random vibration signal $s(t)$ will be written as [35]:

$$\mathrm{WVD}_x(t,f)=\int_{-\infty}^{\infty} x\left(t+\frac{\tau}{2}\right)x^*\left(t-\frac{\tau}{2}\right)e^{-j2\pi f\tau}d\tau \quad (4)$$

where $x(t)$ is the analytic form of $s(t)$. '*' denotes conjugation. τ is the delay of time t. WVD also has its own defect. Cross interference will occur when there are multiple components in the signal. For example, when $x(t)=x_1(t)+x_2(t)$, namely:

$$\begin{aligned}\mathrm{WVD}_x(t,f)=&\mathrm{WVD}_{x_1}(t,f)+\mathrm{WVD}_{x_2}(t,f)\\&+2\mathrm{Re}\{\mathrm{WVD}_{x_1+x_2}(t,f)\}\end{aligned} \quad (5)$$

the cross-interference term is the oscillation frequency interference without physical meaning, which provides false energy distribution, so it needs to be avoided as far as possible.

The idea of SPWVD is to eliminate and suppress the cross interference by windowing in the time domain and frequency domain. Although this operation may lose some resolution of WVD, SPWVD can choose the width of time and frequency window function independently, so it is easy to control the smoothing process. By smoothing the frequency domain of equation (4), we can get:

$$\mathrm{PWVD}_x(t,f)=\int_{-\infty}^{\infty} h(\tau)x\left(t+\frac{\tau}{2}\right)x^*\left(t-\frac{\tau}{2}\right)e^{-j2\pi f\tau}d\tau \quad (6)$$

where $h(\tau)$ is the time window function, which can suppress the cross-interference term of WVD in frequency domain. SPWVD can be obtained by smoothing equation (6) in time domain, namely:

$$\begin{aligned}\mathrm{SPWVD}_x(t,f)=&\int_{-\infty}^{\infty}h(\tau)\int_{-\infty}^{\infty}g(u-\tau)x\left(t+\frac{\tau}{2}\right)x^*\\&\times\left(t-\frac{\tau}{2}\right)e^{-j2\pi f\tau}d\tau du\end{aligned} \quad (7)$$

where $g(u-\tau)$ is the frequency window. It can suppress the cross-interference term of WVD in time domain. Kaiser window function is used in this paper, which can be expressed by the following formula,

$$\phi(\omega)=\frac{I_0\left(\beta\sqrt{1-\left(\frac{2n}{N-1}-1\right)^2}\right)}{I_0(\beta)},0\leqslant n\leqslant N-1 \quad (8)$$

where I_0 is the first zero-order Bessel function. β is the parameter used to adjust the shape of the window, the value of β is 20 in this study. In the discussion section, we also compared the effect of other window functions in SPWVD. The results show that Kaiser window function is better.

To illustrate the effect of SPWVD in suppressing cross-interference, a section of bearing outer race defect signal is selected for analysis. The width of localized defect of bearing outer race is 1.4 mm. See section 5 for a detailed description of the test bearings. Figure 3 plots the time domain of the damaged bearing, from which we can see the periodic fault impulses. WVD is performed on figure 3. The obtained TFD is shown in figure 4, from which there are obvious cross-interference terms between each impulse (in the red dotted line box). In figure 5, SPWVD is used to eliminate the cross-interference terms, and clear TFD results of four impulses are obtained. This is important in the following multi-impulse calculation process. Because multi-impulse calculation can make the localized defect size estimation method automatic and fast in the future.

4. Quantitative estimation of bearing localized defect size

To accurately capture the time nodes of **A** and **C** in figure 2, this paper will determine the ΔT through the variation of instantaneous energy based on the SPWVD time-frequency analysis. Taking the vibration signal in figure 3 as an example, one of the impulses is taken for analysis, as shown in figure 6. In figure 6(a), the low frequency response (entry phase), the high-frequency response (impact phase), and the low frequency attenuation (exit phase) are apparent the time domain. Figure 6(b) is the energy distribution processed by SPWVD. On the right side of figure 6(b) is the colours label, from dark

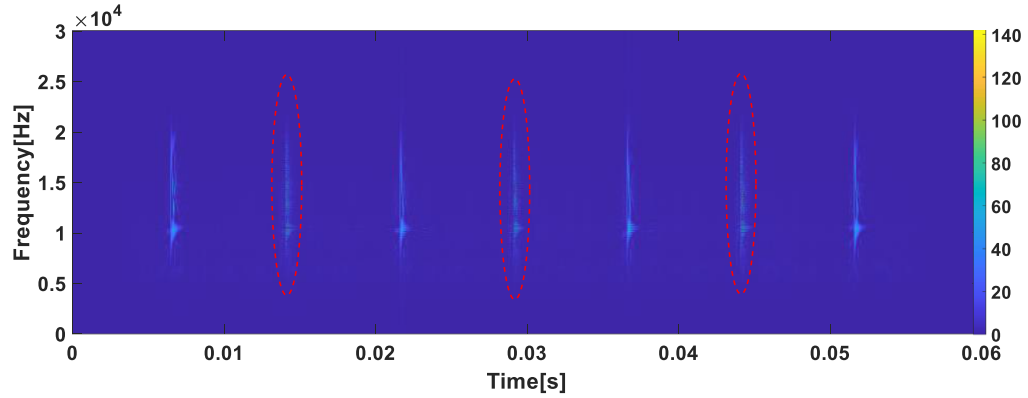


Figure 4. WVD time frequency diagram of figure 3.

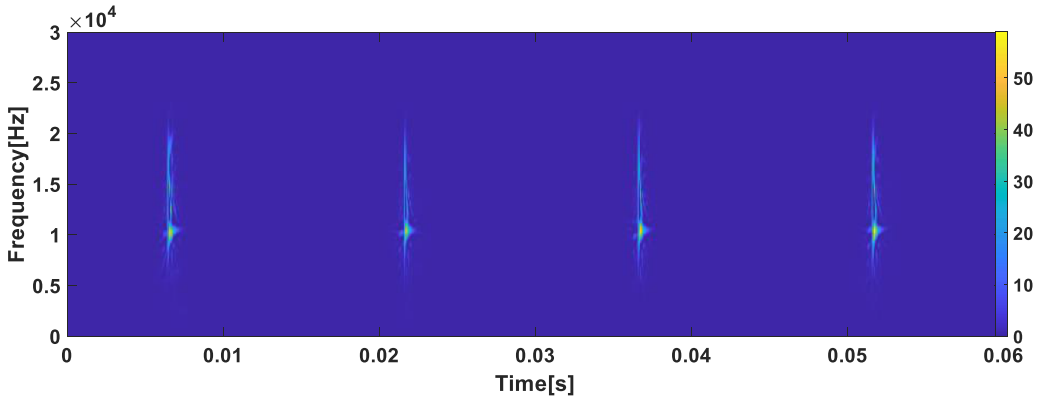


Figure 5. SPWVD time frequency diagram of figure 3.

blue (0) to yellow (35). The colours represent the size of the 'energy density', indicating the amplitude of the energy density at a certain time and a certain frequency. It shows that the vibration energy is concentrated in the impact (high frequency) phase, and the energy is the largest at this time. The energy is the smallest at the phase of entry (low frequency), then, in the exit phase, the energy value of low frequency oscillation is between the collision phase and the entry phase. In addition, the energy when the ball has not entered the defect or has left the defect is so small as to be ignored.

From the above analysis, the time interval ΔT can be estimated by extracting the time nodes of energy variation of the entry phase and the exit phase. However, it will produce a high error to extract time nodes directly from the time domain or two-dimensional SPWVD time-frequency diagram. Therefore, the instantaneous energy $|s(t)|^2$ of the signal on one-dimensional time coordinate can be calculated according to the marginal distribution of SPWVD. The key time nodes of entering and leaving the defect can be extracted by the variation of instantaneous energy on the time axis. $|s(t)|^2$ can be obtained from equation (9), that is:

$$E = |s(t)|^2 = \int_{f_1}^{f_2} SPW_x(t,f)df. \tag{9}$$

The marginal distribution on the time axis of figure 6(b) is presented in figure 7. From picture the instantaneous energy is almost zero at the phase during the ball has not entered the defect area or has left the damage area, which indicates that the vibration frequency has no change at these stages. At **A** point, the instantaneous energy has a small increase, which indicates that the ball is about to enter the damage area, and the contact stress between the ball and the raceways begins to decrease, resulting in a low-frequency response. In the process of collision, it will produce a high-frequency response, accompanied by huge energy, so point **B** is the beginning of the impact phase. When the ball is about to leave the fault area, due to its re-contact with the raceways, the re-stress process will generate a low-frequency response again. After point **C**, the rolling element leaves the damage area completely, accompanied by a low-frequency attenuation process. Therefore, the time interval of the rolling element traversing over the damage will be determined by the ΔT between point **A** and point **C**. The relationship between ΔT and bearing localized damage size l on the outer race can be expressed by equation (10), according to figure 8, where $D = D_p + d$ (mm), D_p is the pitch diameter. d is the diameter of ball.

$$l = 2\pi \cdot \frac{D}{2} \cdot f_c \cdot \Delta T = \pi D f_c \Delta T \tag{10}$$

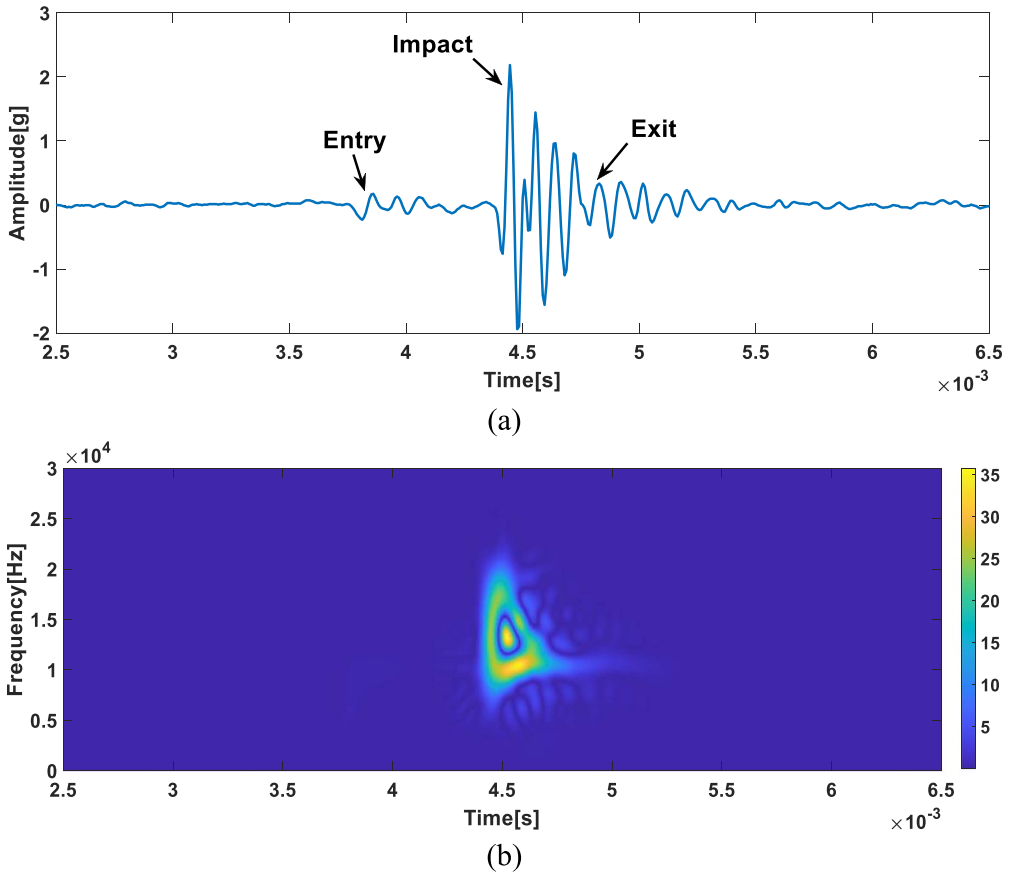


Figure 6. (a) The single impulse of figure 3. (b) SPWVD time-frequency diagram of (a).

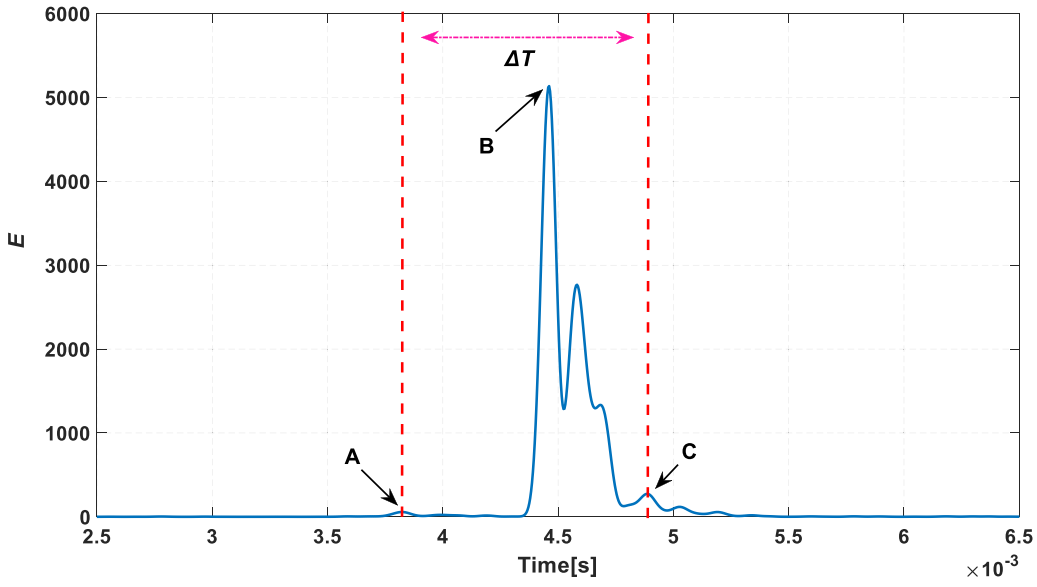


Figure 7. The instantaneous energy curve of the single impulse response.

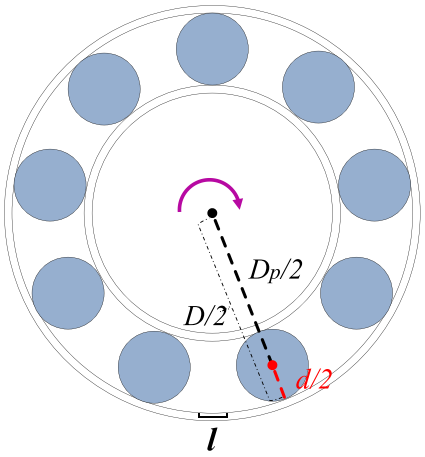


Figure 8. Schematic diagram of ball bearing (localized damage on outer race).

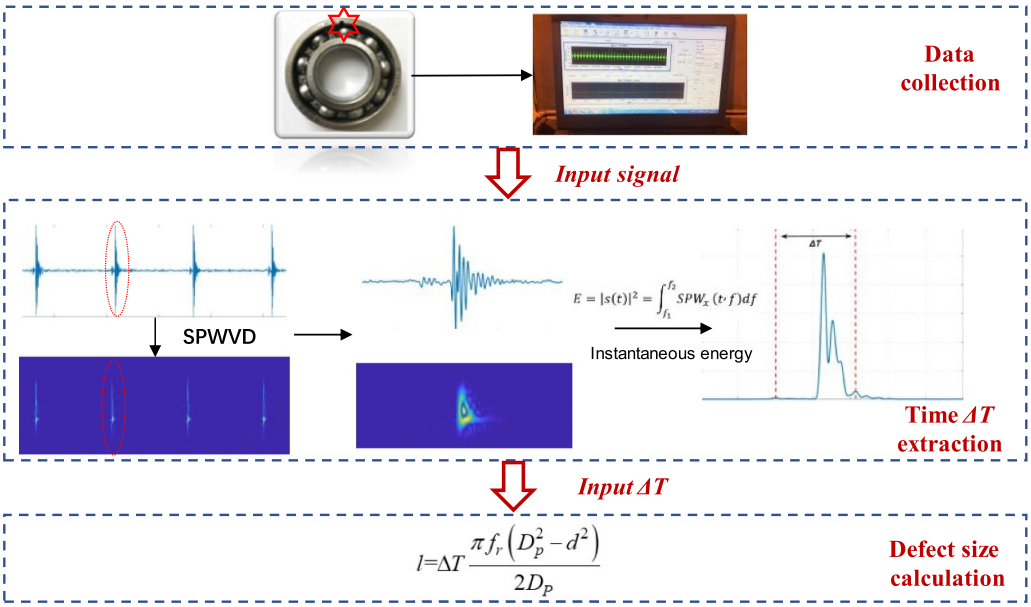


Figure 9. The process of calculating localized defects of ball bearings.

where f_c is the cage frequency (Hz). The relationship between f_c and the rotation frequency f_r is as follows:

$$f_c = \frac{f_r}{2}\left(1 - \frac{d}{D_p}\cos(\theta)\right) \tag{11}$$

where θ is the contact angle, the θ of the deep groove ball bearings in this paper is equal to 0. So, it can be obtained:

$$l = \Delta T \frac{\pi f_r \left(D_p^2 - d^2\right)}{2D_P}. \tag{12}$$

As for inner race defect, the result is the same as equation (12). The specific formula derivation can be found in [16]. Figure 9 is a flow chart for calculating localized defects of ball bearings in this paper.

5. Experimental verification

5.1. Test equipment

To fully verify the practicability of the proposed method, experiments of different localized defect sizes on bearings' raceways were carried out in this section. Different defects made by wire-electrode cutting were planted on the outer race as well as inner race, respectively, whose widths (circumferential length) include 1.0, 1.4, 1.8, and 2.2 mm, and all of the defect depths are 1.0 mm. HRB 6206 ball bearings were the test object, as shown in figure 10. Table 1 gives the dimensions of the bearing. The tests relied on the ABLT-1A Bearing Testing Machine which one fault bearing was placed into the operating bin, accompanied by three normal bearings. AI002 vibration acceleration sensors were installed on each bearing, as shown in figure 11. To ensure the quality of the signals, the sampling frequency was 128 kHz. JM5938 was used as data

Figure 10. The faulty bearings with localized defects on raceways.

Table 1. HRB6206 ball bearing parameters.

Type	Pitch diameter (mm)	Diameter of ball (mm)	Ball number	Contact angle
6206	46	9.5	9	0

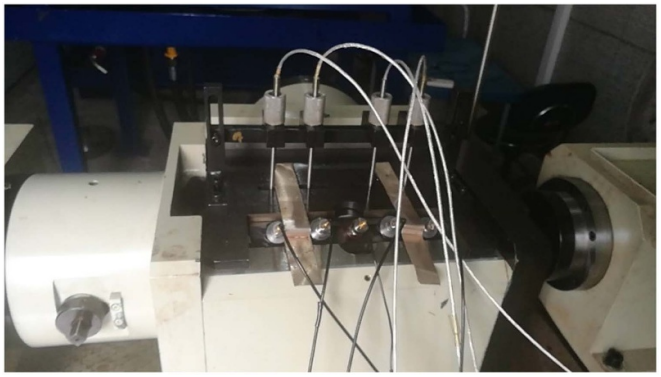


Figure 11. ABLT-1A bearing testing machine.

acquisition. The bearings were driven by a motor spindle and the test speeds were 1200, 1500, 1800, and 2400 rpm.

5.2. Test results

Take the inner race localized defects as the examples, figure 12 draws the time domains when a ball passing through the localized defect under four different sizes at 1200 rpm. From the pictures, we can see that the time interval between entering phase and collision is increasing. However, it is hard to accurately pick up the leaving moment directly from the fault impulses. Figure 13 plots the variation of instantaneous energy of four fault impulses. As can be seen from the results that the variation law of instantaneous energy is identical for four situations. When the ball is on the phase of entry, the instantaneous energy will fluctuate upward due to the low-frequency response. At the impact phase, the high-frequency response will make the instantaneous energy increase sharply and oscillate. Then, the ball leaves the defect and re-contacts with the raceways, and the low-frequency response produced by this process will cause an instantaneous energy peak again. After that, the energy oscillation is gradually zero. It is worth noting that the instantaneous energy variation caused by the low-frequency response is smaller than that caused by the high-frequency response, and the variation caused by the re-stress process is larger than that caused by the de-stress process. Besides, in other phases, the instantaneous energy is almost zero, with no obvious variation. Therefore, according to the time interval ΔT between the two low-frequency responses, the damage size of the inner race can be calculated whose results are basically consistent with the actual sizes.

To ensure the accuracy of the estimation sizes, five impulses are selected for calculation under the case of each defect size, and the average values and errors are given in tables 2 and 3, respectively. We can observe from the tables that the evaluation sizes of the proposed method are very accurate and the errors are less than 5%, whether in the calculation of outer race damage or inner race damage.

5.3. Comparison with wavelet time frequency distribution

In the aspect of TFD, linear TFD is also a common method and is extensively used in the field of signal processing, in which wavelet analysis has high resolution, especially in the low-frequency region. Thus, this paper selects CWT to analyze the fault impulses of bearing localized defects and compares them with the proposed method. CWT can be expressed as [36]:

$$\mathrm{CWT}(\alpha,\tau) = \frac{1}{\sqrt{\alpha}}\int x(t)\,\psi^*\left(\frac{t-\tau}{\alpha}\right)dt \tag{13}$$

where α denotes the scale, τ is the time translation. $\psi^*(\bullet)$ is the complex conjugate of the basic wavelet or mother wavelet $\psi(\bullet)$. Complex Morlet wavelet have variable bandwidth σ and centre frequency f_0, besides its shape is very similar to the impact signal, which is conducive to the analysis of the bearing impact impulse. Therefore, complex Morlet wavelet is chosen as the basic wavelet function, which is written as:

$$\psi(t) = \frac{\sigma}{\sqrt{\pi}} e^{-\sigma^2 t^2} e^{i2\pi f_0 t}. \tag{14}$$

Due to space limitations, figures 14 and 15 only draw the results when the localized defect size of the outer race and inner race is 1.0 and 2.2 mm, respectively. From the figures, we can see the resolution of CWT is poor compared with SPWVD in extracting low-frequency time nodes. Especially in the exit phase, it is unable to accurately extract the time separation nodes of collision phase and exit phase, which makes the estimation results and the actual results produce large errors. Certainly, parameters σ and f_0 will affect the resolution of CWT. The influence of these two parameters will be discussed in section 6.

5.4. Comparison with Sawalhi's method

To further illustrate the accuracy of the proposed method. Sawalhi's method [16] is selected for comparison. Sawalhi extracted the time nodes named time to impact of the entry

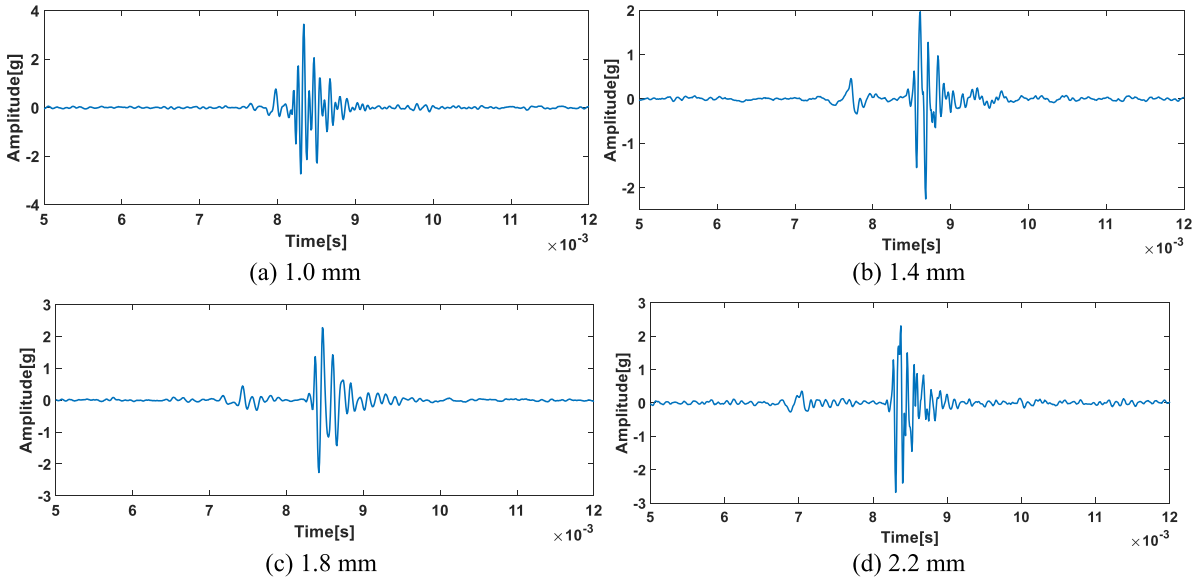


Figure 12. Time domain waveforms of different localized defect sizes on inner race.

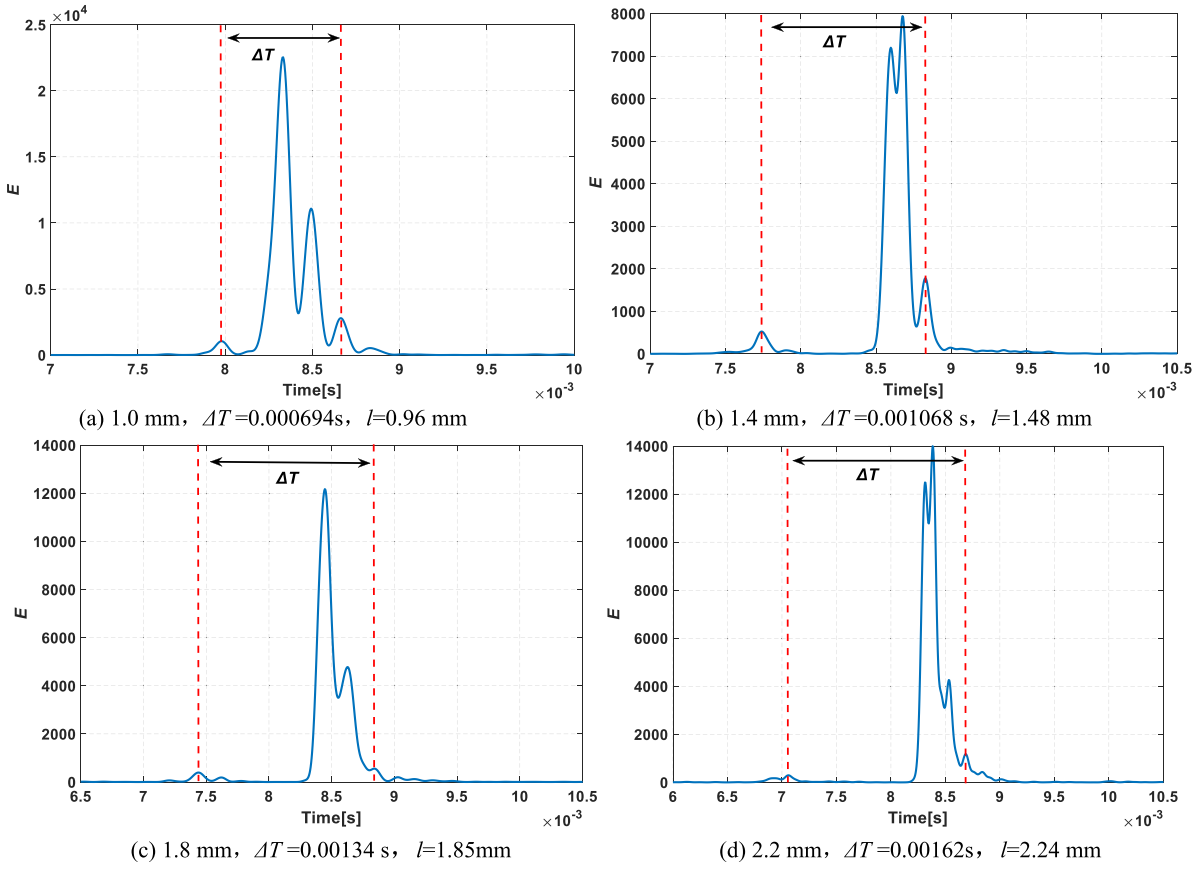


Figure 13. Instantaneous energy variation of different localized defect sizes on inner race.

phase (**A** point) and impact phase (**B** point). Firstly, AR filtering is applied to the signal, and then the signal filtered by using complex Morlet wavelet analysis to enhance these two features. Finally, the time interval between A and B is extracted by the square envelope method. The method flow is plotted in figure 16. Significantly, Sawalhi thinks that from **A** to **B** is half the time for the rolling element traversing over the damage size. Therefore, it is necessary to multiply the extracted time by two.

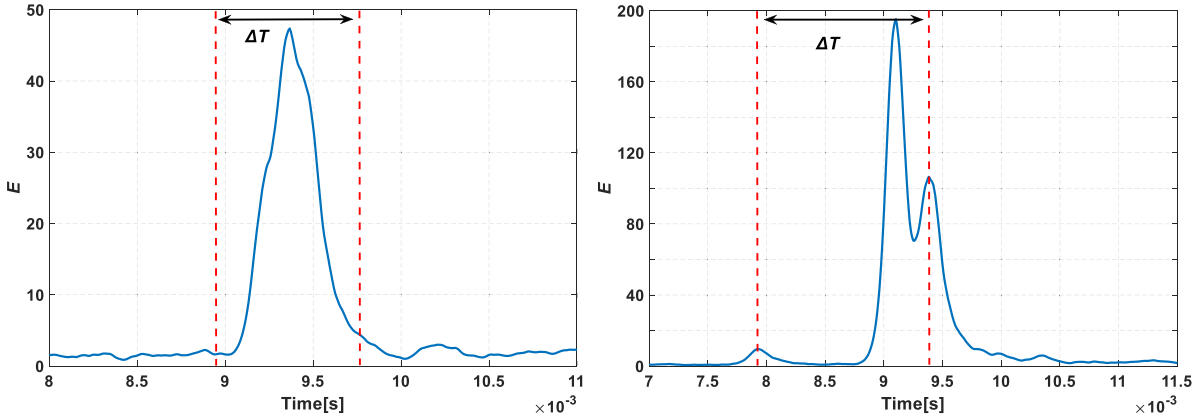
Figure 17 draws the comparison between the proposed method and Sawalhi's method in estimating four defect sizes of the outer race. The results of the inner race are shown in

Table 2. Estimation results and errors of different defect sizes of outer race (1200 rpm).

Defect size (mm)	Impulse 1	Impulse 2	Impulse 3	Impulse 4	Impulse 5	Mean value	Error
1.0	1.07	0.92	1.04	0.90	0.97	0.98	2.0%
1.4	1.46	1.47	1.45	1.47	1.43	1.46	4.3%
1.8	1.83	1.78	1.85	1.88	1.77	1.82	1.1%
2.2	2.11	2.17	2.14	2.20	2.17	2.16	1.9%

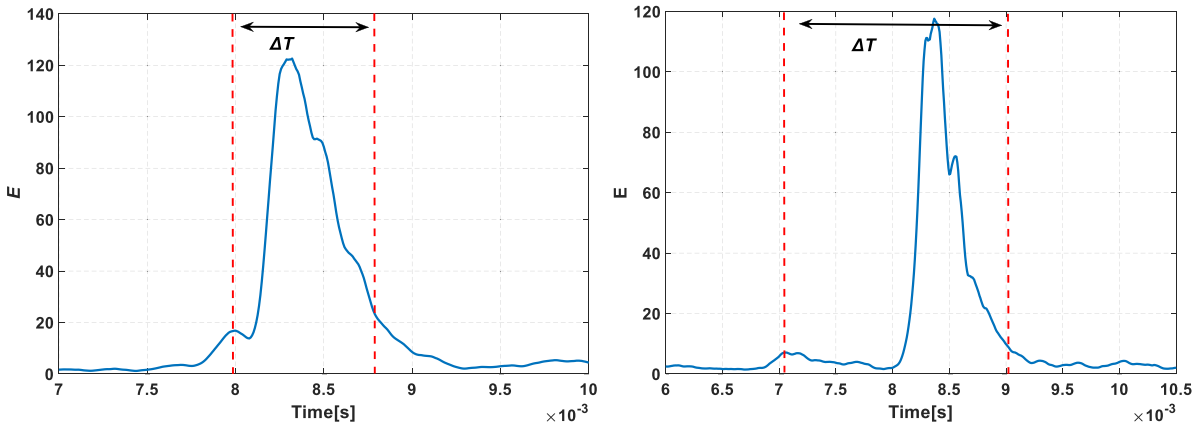
Table 3. Estimation results and errors of different defect sizes of inner race (1200 rpm).

Defect size (mm)	Impulse 1	Impulse 2	Impulse 3	Impulse 4	Impulse 5	Mean value	Error
1.0	0.97	1.06	0.93	0.94	0.92	0.96	4.0%
1.4	1.48	1.53	1.44	1.42	1.35	1.44	2.8%
1.8	1.85	1.79	1.88	1.74	1.87	1.83	1.4%
2.2	2.24	2.17	2.14	2.11	2.23	2.18	1.0%



(a) 1.0 mm，ΔT =0.000831 s，l= 1.15 mm (b) 2.2 mm，ΔT = 0.001461s，l=2.02 mm

Figure 14. CWT instantaneous energy analysis of outer race localized defect size.



(a) 1.0 mm，ΔT = 0.000805s，l= 1.11 mm (b) 2.2 mm，ΔT =0.001977 s，l=2.73 mm

Figure 15. CWT instantaneous energy analysis of inner race localized defect size.

figure 18. To make it fair, each estimation method selects five identical fault impulses to get the mean value of the calculation results. From the pictures, we can see that the error of Sawalhi's method is relatively small when the defect size is 1.0 mm. However, with the increase of the damage width, the error of Sawalhi's method will grow. The limitation of Sawalhi's method will be detailed in the discussion section. On the contrary, the proposed method base on SPWVD is more precise and in line with the actual size whether in the outer race defect or inner race defect.

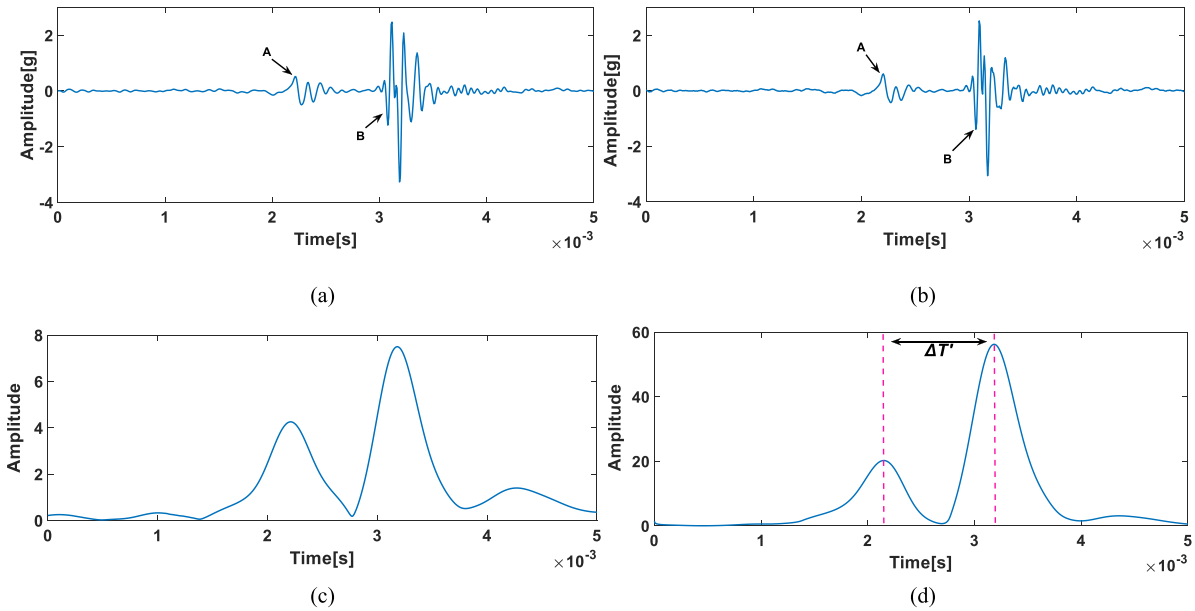


Figure 16. Sawalhi's method: (a) vibration signal; (b) AR filtering signal; (c) after complex Morlet wavelet filtered; (d) squared enveloped signal.

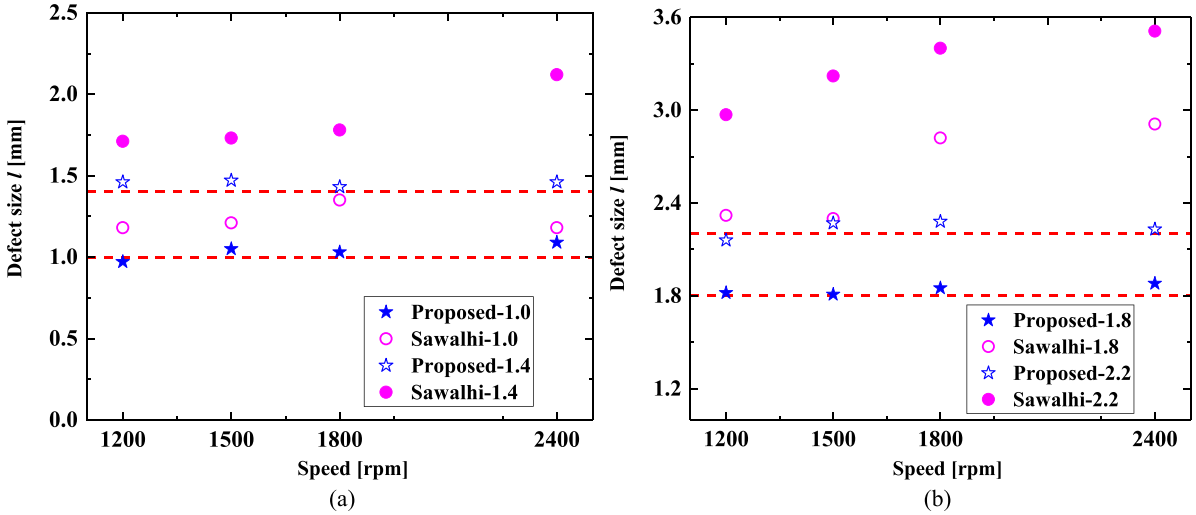


Figure 17. Comparison of four defect sizes estimation results of outer race at different speed. (a) 1.0 mm and 1.4 mm. (b) 1.8 mm and 2.2 mm.

6. Discussions

As mentioned in section 5.3, bandwidth σ and centre frequency f_0 of the complex Morlet wavelet will affect the resolution of the impact signal's TFD. Take the inner race localized defect sizes of 1.0 mm (figure 12(a)) and 2.2 mm (figure 12(d)) as examples. Figures 19 and 20 analyse the variation of instantaneous energy of fault impulse under different centre frequencies and bandwidths of complex Morlet wavelet, respectively. The results show that with the increase of centre frequency (or bandwidth), the instantaneous energy curve will be smoother. There will be less low-frequency detail of the ball's exit phase. Although the corresponding time node of exit low frequency can be found in the smaller centre frequency (or bandwidth), there are too many interference peaks in the curve, which is easy to cause misjudgement. Therefore, CWT is inferior to SPWVD in the resolution of extracting instantaneous energy.

For the same reason, in SPWVD, a variety of window functions can be selected to smooth the signal. Figure 21 draws the instantaneous energy curves of figures 12(a) and (d) under five-window functions. The five-window functions are Kaiser, Hanning, Hamming, Triangle, and Blackman. From the result, we can know that the instantaneous energy curves of five window functions are higher than CWT in the low-frequency resolution. Since the Kaiser window curve in figure 21(b) has less interference at other frequencies, the Kaiser window is used for the result estimation in this paper.

In [16], Sawalhi regarded time to impact (TTI) as half ΔT (the time of the rolling element passing through the damage size), and some studies [17–19] are also based on this

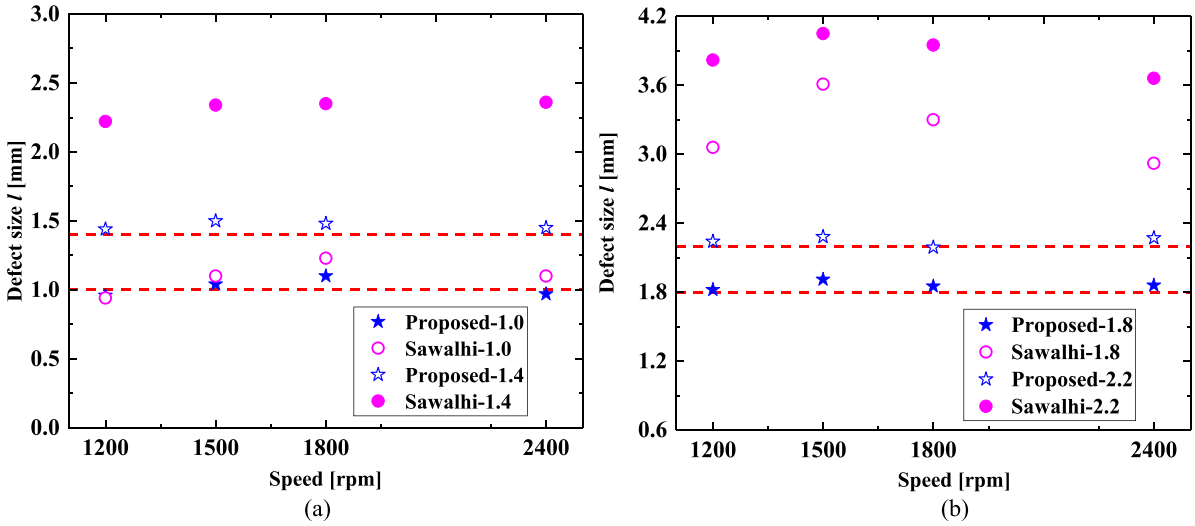


Figure 18. Comparison of four defect sizes estimation results of inner race at different speed. (a) 1.0 mm and 1.4 mm. (b) 1.8 mm and 2.2 mm.

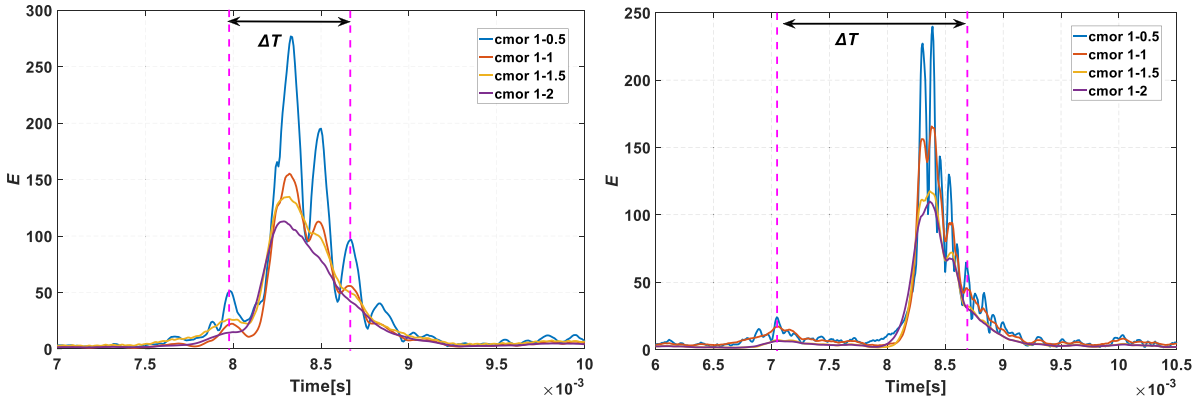


Figure 19. The results of instantaneous energy at different complex Morlet wavelet centre frequencies.

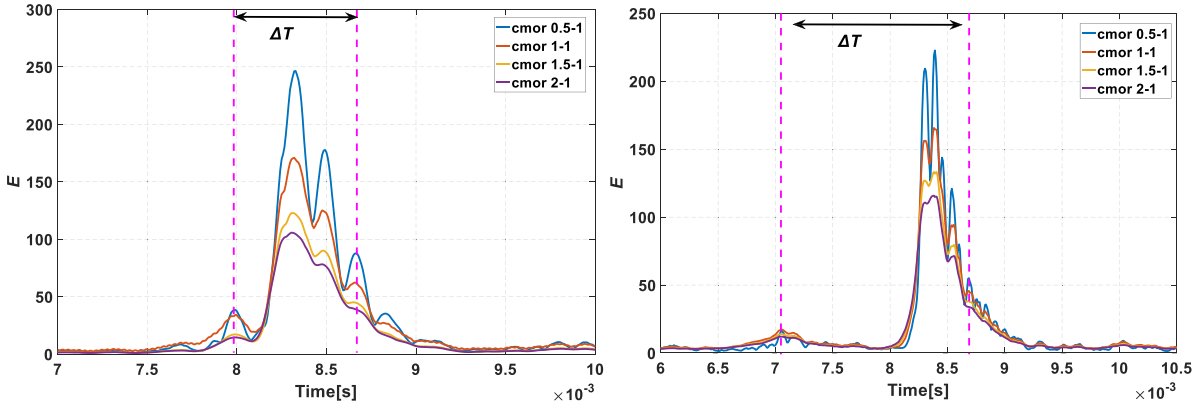
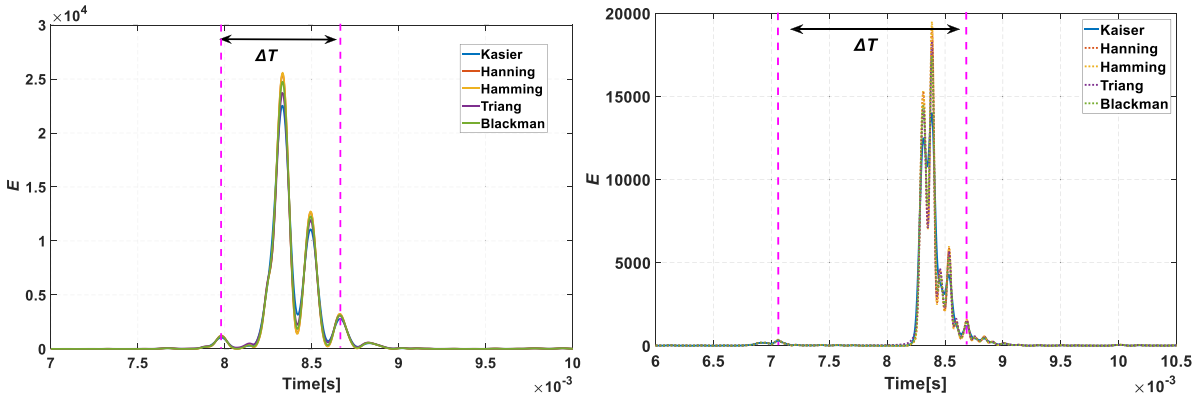


Figure 20. The results of instantaneous energy at different complex Morlet wavelet bandwidths.

theory. However, this method has limitations. We can find from figures 17 and 18 that Sawalhi's method is relatively accurate in the case of small size but when the defect size becomes larger, the estimated size is larger. A reasonable explanation will be given below. Assuming that the rolling element only has a major impact on the trailing edge of the damage in the ideal case, it can be roughly divided into three cases as shown in figure 22.



(a) 1.0 mm (b) 2.2 mm

Figure 21. Instantaneous energy of SPWVD under different five window functions.

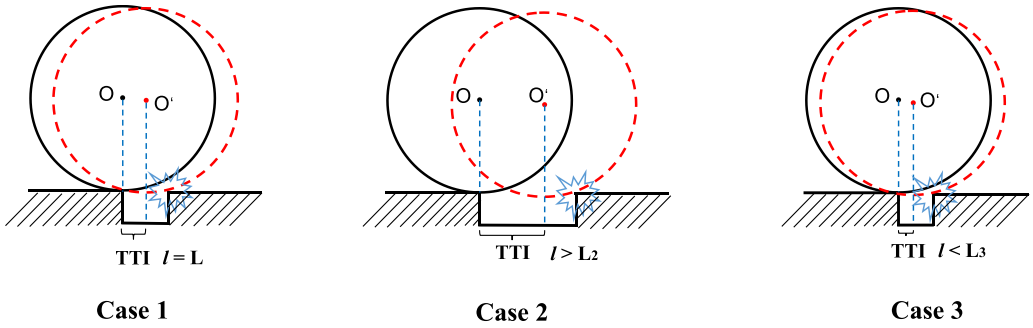


Figure 22. The process of impact between rolling element and defect trailing edge under ideal condition.

Case 1. TTI is half ΔT. The estimated size $\boldsymbol{l}$ will be consistent with the actual size $\boldsymbol{L}$.

Case 2. If the actual size $\boldsymbol{L_2}$ is larger than $\boldsymbol{L}$, the TTI will be larger than $\Delta T/2$, and the estimated size $\boldsymbol{l}$ will be larger than the actual size $\boldsymbol{L_2}$.

Case 3. If the actual size $\boldsymbol{L_3}$ is smaller than $\boldsymbol{L}$, the TTI will be smaller than $\Delta T/2$, and the estimated size $\boldsymbol{l}$ will be smaller than the actual size $\boldsymbol{L_3}$.

Therefore, Sawalhi's method is suitable for the estimation of small size defects because the estimation error is relatively small. The premise is that the fault size is very small relative to the diameter of the rolling element. Another reason for the error is that the uncertainty of the impacts should not be ignored in reality. As we know, inner race fault will produce modulation phenomenon. The defect position will move with the rotation cycle, which will increase the uncertainty of the impact process. Figure 23 shows the fault impulses with inner race defect size of 1.4 mm at 1200 rpm, where the location of the maximum impact is different. In the impact phase, it is difficult to determine which moment is the main impact because the collisions between the upper edge of the rolling element and the raceway may be more intense. When Sawalhi's method is used, AR filtering and wavelet analysis tend to improve the maximum impact amplitude and the square envelope prefers the maximum amplitude as the main peak, which will also make the result of TTI larger than the actual size. Therefore, the authors think that the impact phase should not be taken as the main analysis time. In the exit phase, the low-frequency oscillation feature of the re-stress process deserves more attention.

Howard *et al* [21–24] were the first to pay attention to the low-frequency oscillation response in the exit phase. They also gave the estimation formula of localized defect size, namely [22]:

$$L = \sin(2\beta_2 + \beta_1) \times R \tag{15}$$

where β_1 is the angular travel which associate with the relative displacement of the inner race and outer race during the phase of entry to impact. β_2 is the angular travel which is related to the maximum total elastic contact deformation of the rolling element in the transient phase of inlet and outlet. R is the radius of the outer raceway, that is $R = (D_P + d)/2$. However, the estimation process of these two parameters is more complex and increases the probability of error occurrence. In addition, only the localized defects of the outer race are discussed. The implementation of some researchers' methods is also complex and it is difficult to get accurate parameters, such as bearing stiffness, maximum load, contact stress [21–23, 27–31] and so on. Therefore, this paper is devoted to give a relatively simple method to quantitatively assessment the localized defect of inner race and outer race through the variation of instantaneous energy. The instantaneous energy only focuses on the relative frequency variation of the rolling element when passing

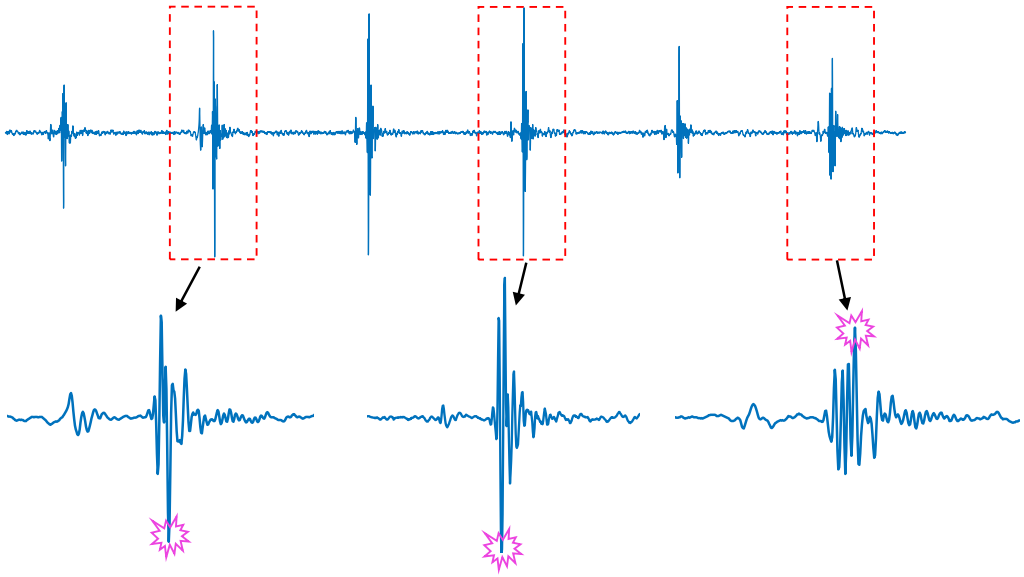


Figure 23. The maximum impact amplitudes caused by inner race modulation are different.

through the defect, and is not affected by the angular travel, amplitude and inherent parameters of the rolling element.

It is worth mentioning that the fault signals in this paper have a high-quality signal-to-noise ratio. There are two main reasons. One is that we adopt a very high sampling frequency in the test. The sampling frequency is 128 kHz. The higher the sampling frequency, the higher the restoration of the signal, which can clearly capture the details of the fault impulse. The second is that the test vibration sensor is close to the bearing position without too many transmission paths, so there is less interference noise. In some actual working conditions, the signal will be mixed with high noise, which requires more signal processing technologies to eliminate noise and enhance useful features. The quantitative calculation of bearing damage under strong noise is also a difficult problem in this field. It is also worthy of attention and in-depth study in the future.

7. Conclusions

This paper presents a convenient and accurate quantitative estimation method of ball bearing localized defect size based on vibration instantaneous energy analysis. The time interval between the entry event and the exit event of the ball is determined by analyzing the time-frequency energy distribution (SPWVD) of the fault impulses. The bearing fault tests at different speeds and different localized defect sizes indicate that the proposed method has high accuracy in the estimation of the defect size of the outer race and inner race. Some conclusions are as follows:

(a) The proposed method can well extract the time nodes of high-frequency and low-frequency responses in the fault impact, especially distinguishing the low-frequency moment of the exit phase, which is vulnerable to the interference of high-frequency responses of the impact phase. The errors between the results of defect size estimation and the actual sizes are small.

(b) CWT has poor accuracy in instantaneous energy estimation compared with the proposed method, which is reflected in the extraction of low-frequency response time node in the exit phase. Moreover, by adjusting the center frequency and bandwidth of the wavelet, the estimation results are also not ideal. Due to the uncertainty of the impact phase, Sawalhi's method will have a larger error in defect size estimation of the outer race and inner race, so more attention should be paid to the low-frequency response at the exit phase in the future.

(c) The vibration instantaneous energy estimation method of bearing localized defect size is easy to implement and it does not need to know those complex bearing inherent parameters in advance, which is more convenient for practical engineering applications.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

This paper is supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX20_0211), National Science and Technology Major Project (J2019-IV-0004-0071), and China Scholarship Council (202106830045).

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

Zhiyuan He: Conceptualization, Methodology, Software, Data Collection and Analysis, Original manuscript Writing, Visualization. Guo Chen: Funding acquisition, Writing—review and editing, Supervision, Project administration. Kaiyong Zhang: Experimental design, Data Collection, Resources. Konstantinos Gryllias: Writing—review and editing, Visualization, Supervision.

ORCID iDs

Zhiyuan He https://orcid.org/0000-0001-8283-6773
Konstantinos Gryllias https://orcid.org/0000-0002-8703-8938

References

[1] Harris T A and Kotzalas M N 2007 *Rolling Bearing Analysis: Essential Concepts of Bearing Technology* 5th edn (New York: Taylor and Francis) (https://doi.org/10.1201/9781420006599)
[2] Ho D and Randall R B 2000 Optimisation of bearing diagnostic techniques using simulated and actual bearing fault signals *Mech. Syst. Signal Process.* **14** 763–88
[3] Yan R, Gao R and Chen X 2014 Wavelets for fault diagnosis of rotary machines: a review with applications *Signal Process.* **96** 1–15
[4] Xu Y, Tian W, Zhang K and Ma C 2019 Application of an enhanced fast kurtogram based on empirical wavelet transform for bearing fault diagnosis *Meas. Sci. Technol.* **30** 035001
[5] Tomasz B and Sawalhi N 2012 Fault detection enhancement in rolling element bearings using the minimum entropy deconvolution *Arch. Acoust.* **37** 131–41
[6] Wang S, Xiang J, Tang H, Liu X and Zhong Y 2019 Minimum entropy deconvolution based on simulation-determined band pass filter to detect faults in axial piston pump bearings *ISA Trans.* **88** 186–98
[7] He Z, Chen G, Hao T, Liu X and Teng C 2021 An optimal filter length selection method for MED based on autocorrelation energy and genetic algorithms *ISA Trans.* **109** 269–87
[8] Li H, Zhang Y and Zheng H 2009 Hilbert-Huang transform and marginal spectrum for detection and diagnosis of localized defects in roller bearings *J. Mech. Sci. Technol.* **23** 291–301
[9] Zhang M, Jiang Z and Feng K 2017 Research on variational mode decomposition in rolling bearings fault diagnosis of the multistage centrifugal pump *Mech. Syst. Signal Process.* **93** 460–93
[10] Cao H, Fan F, Zhou K and He Z 2016 Wheel-bearing fault diagnosis of trains using empirical wavelet transform *Measurement* **82** 439–49
[11] Zhao X and Jia M 2018 A novel deep fuzzy clustering neural network model and its application in rolling bearing fault recognition *Meas. Sci. Technol.* **29** 125005
[12] Li C, Li S, Zhang A, He Q, Liao Z and Hu J 2021 Meta-learning for few-shot bearing fault diagnosis under complex working conditions *Neurocomputing* **439** 197–211
[13] Lei Y, Li N, Guo L, Li N, Yan T and Lin J 2018 Machinery health prognostics: a systematic review from data acquisition to RUL prediction *Mech. Syst. Signal Process.* **104** 799–834
[14] Cerrada M, Sánchez R V, Li C, Pacheco F, Cabrera D, Oliveira J V and Vásquez R E 2018 A review on data-driven fault severity assessment in rolling bearings *Mech. Syst. Signal Process.* **99** 169–96
[15] Epps I K 1991 An investigation into vibrations excited by discrete faults in rolling element bearings *PhD Thesis* School of Mechanical Engineering, The University of Canterbury, Christchurch, New Zealand (https://doi.org/10.26021/1241)
[16] Sawalhi N and Randall R B 2011 Vibration response of spalled rolling element bearings: observations, simulations and signal processing techniques to track the spall size *Mech. Syst. Signal Process.* **25** 846–70
[17] Sawalhi N, Wang W and Becker A 2018 Vibration signal processing using cepstrum editing technique to enhance spall-related vibration features in rolling element bearings *Int. J. Mech. Eng. Robot.* **8** 65–68
[18] Wang W, Sawalhi N and Becker A 2016 Size estimation for naturally occurring bearing faults using synchronous averaging of vibration signals *J. Vib. Acoust.* **138** 051015
[19] Zhao S, Liang L, Xu G, Wang J and Zhang W 2013 Quantitative diagnosis of a spall-like fault of a rolling element bearing by empirical mode decomposition and the approximate entropy method *Mech. Syst. Signal Process.* **40** 154–77
[20] Chen A and Kurfess T R 2019 Signal processing techniques for rolling element bearing spall size estimation *Mech. Syst. Signal Process.* **117** 16–32
[21] Petersen D, Howard C Q and Prime Z 2015 Varying stiffness and load distributions in defective ball bearings: analytical formulation and application to defect size estimation *J. Sound Vib.* **337** 284–300
[22] Moazen-Ahmadi A, Howard C Q and Petersen D 2016 The path of rolling elements in defective bearings: observations, analysis and methods to estimate spall size *J. Sound Vib.* **366** 277–92
[23] Moazen-Ahmadi A and Howard C Q 2016 A defect size estimation method based on operational speed and path of rolling elements in defective bearings *J. Sound Vib.* **385** 138–48
[24] Larizza F, Howard C Q and Grainger S 2021 Defect size estimation in rolling element bearings with angled leading and trailing edges *Struct. Health Monit.* **30** 1102–16
[25] Zhang H, Borghesani P, Smith W A, Randall R B, Shahriar M R and Peng Z 2021 Tracking the natural evolution of bearing spall size using cyclic natural frequency perturbations in vibration signals *Mech. Syst. Signal Process.* **151** 107376
[26] Zhang H, Borghesani P, Randall R B and Peng Z 2022 A benchmark of measurement approaches to track the natural evolution of spall severity in rolling element bearings *Mech. Syst. Signal Process.* **166** 108466
[27] Chen A and Kurfess T R 2018 A new model for rolling element bearing defect size estimation *Measurement* **114** 144–9
[28] Kogan G, Bortman J and Klein R 2017 A new model for spall-rolling-element interaction *Nonlinear Dyn.* **87** 219–36
[29] Kogan G, Klein R and Bortman J 2018 A physics-based algorithm for the estimation of bearing spall width using vibrations *Mech. Syst. Signal Process.* **104** 398–414
[30] Luo M, Guo Y, Wu X and Na J 2019 An analytical model for estimating spalled zone size of rolling element bearing

based on dual-impulse time separation *J. Sound Vib.* **453** 87–102
[31] Patil A P, Mishra B K and Harsha S P 2021 A mechanics and signal processing based approach for estimating the size of spall in rolling element bearing *Eur. J. Mech.* A **85** 104125
[32] Tang X, Guo Y, Ding Y and Zheng H 2012 Application of rolling element bearing envelope analysis based on short time Fourier transition and independent components analysis *J. Mech. Strength* **34** 1–5 (in Chinese)
[33] Cheng J S, Yu D J and Yang Y 2007 Application of an impulse response wavelet to fault diagnosis of rolling bearings *Mech. Syst. Signal Process.* **21** 920–9
[34] Bin G F, Liao C J and Li X J 2011 The method of fault feature extraction from acoustic emission signals using Wigner-Ville distribution *Adv. Mater. Res.* **216** 732–7
[35] Cohen I 1995 *Time-Frequency Analysis: Theory and Applications* (Englewood Cliffs, NJ: Prentice-Hall)
[36] Rioul O and Vetterli M 1991 Wavelets and signal processing *IEEE Signal Proc. Mag.* **8** 14–38