Methods for Rolling Element Bearing Fault Diagnosis under Constant and Time-varying Rotational Speed Conditions

by

Huan Huang

A Dissertation Submitted to the Faculty of Engineering in Fulfillment of the Requirement for the Degree of

Doctorate of Philosophy

in

Mechanical Engineering

Ottawa-Carleton Institute for Mechanical and Aerospace Engineering

University of Ottawa

Abstract

Bearings are among the most commonly used components in rotating machines and bearing fault diagnosis has been widely investigated, especially with vibration signal based techniques. Generally, bearing faults can be diagnosed in the frequency domain since each type of fault has a specific Fault Characteristic Frequency (FCF) proportional to the rotational frequency. However, this approach is complicated by the facts that: (1) bearing vibration signals are often contaminated by random noise and interference signals transmitted from other sources, (2) bearings often operate under time-varying speed conditions which make the FCF also time-varying.

To address the problems for bearing fault signature extraction, new methods are proposed in this thesis. For the constant speed case, bearing fault signature extraction using the method of Oscillatory Behavior-based Signal Decomposition (OBSD) is investigated, which includes: (1) effects of parameter selection and (2) automatic parameter selection of OBSD for bearing fault signature extraction. For bearing fault diagnosis under time-varying speed conditions, research based on the time-frequency technique is conducted, including (1) a multiple time-frequency curve extraction algorithm and (2) a resampling-free and tachometer-free method for the time-varying speed case with the presence of interferences, and (3) proposing a short-time kurtogram for bearing fault signature extraction under time-varying speed conditions.

The effectiveness of the proposed methods in this thesis has been validated by simulated signals and experimental data.

Acknowledgements

First of all, I would like to sincerely thank my supervisor, Dr. Ming Liang, for his patient guidance throughout my Ph.D. study. His immense knowledge, professional suggestions and comments have been essential for this thesis. His great passion for research also motivated me to complete this thesis.

I would also express my special gratitude to my co-supervisor, Dr. Natalie Baddour, who encouraged me by all means. I appreciate her willingness to manage weekly meetings for me on her busy schedule and provide immense help for research and writing. She has also made an effort to strengthen my self-confidence. This thesis would not have been possible without her continuous support.

My sincere thanks also goes to Dr. Dan Necsulescu, Dr. Patrick Dumond, Dr. Jie Liu of Carleton University, and Dr. Youmin Zhang of Concordia University for being my thesis committee members. Their insightful comments helped improve the thesis and their hard questions motivated me to widen my research.

I am grateful to my colleagues in Professor Liang's research group for their friendly help in the lab.

I would like to thank Su, my boyfriend, for his companionship all the way through. Last but not least, I must express my gratitude to my parents for their unconditional love and support.

Table of Contents

List of Figures

List of Tables

Acronyms

ADMM	Alternating Direction Method of Multipliers
ALE	Adaptive Line Enhancement
ANC	Adaptive Noise Cancellation
BPFI	Ball-Pass Frequency of Inner-race
BPFO	Ball-pass Frequency of Outer-race
CI	Convergence Index
DFT	Discrete FourierTransform
DWT	Discrete Wavelet Transform
EEMD	Ensemble Empirical Mode Decomposition
FCC	Fault Characteristic Coefficient
FCF	Fault Characteristic Frequency
GD	Generalized Demodulation
HFRT	High-Frequency Resonance Technique
HPS	High-Pass Scaling
IF	Instantaneous Frequency
IFCF	Instantaneous Fault Characteristic Frequency
IGD	Iterative Generalized Demodulation
ISRF	Instantaneous Shaft Rotational Frequency
LMS	Least-Mean-Square
LPS	Low-Pass Scaling
MCA	Morphological Component Analysis
MTFCE	Multiple Time-Frequency Curve Extraction

OBSD	Oscillatory Behavior-based Signal Decomposition
OCD	Over-Complete Dictionary
RLS	Recursive Least-Square
SALSA	Split Augmented Lagrangian Shrinkage Algorithm
SANC	Self-Adaptive Noise Cancellation
SI	Smoothness Index
SK	Spectral Kurtosis
STFT	Short-Time Fourier Transform
TFR	Time-Frequency Representation
TKR	Time-Kurtogram Representation
TQWT	Tunable Q-factor Wavelet Transform
WT	Wavelet Transform

Chapter 1 Introduction

1.1 Basic concepts

Condition monitoring can significantly reduce the cost of maintenance and can help to avoid accidents caused by machine failure. Rotating machines such as pumps, compressors, gas turbines, steam turbines and electric motors, play crucial roles in many industries. Therefore, condition monitoring of rotating machinery is an area in need of further development [1]. It is reported that 99% of mechanical failures are preceded by noticeable indicators [2]. Thus, a fast and reliable condition monitoring system relies on the detection and diagnosis of incipient faults.

Bearings are among the most frequently used components in rotating machines. The health condition of rotating machines is strongly affected by bearing faults. For instance, according to studies conducted by the Electric Power Research Institute (EPRI), IEEE Industry and General Applications (IEEE-IGA) and B&K company, about 40% of failures in motors are caused by bearing failure [3]. It follows that bearing fault detection is important and has been intensively investigated over the past few decades.

Techniques using different types of physical information have been explored for bearing fault detection and condition monitoring, such as thermal imaging ([4]–[6]), vibration analysis ([7]–[10]), acoustic emissions ([11]–[13]) and lubricant analysis [14]–[16]. Among these techniques, analysis of vibration signals in the time and frequency domains is the most widely used ([17], [18]). Statistical approaches using the root-mean-square, skewness, kurtosis and normalized higher order central moments of vibration signals in the time domain are well-known for distinguishing normal and defective bearings ([19]–[21]). However, statistical features are not sufficient for bearing fault diagnosis since they cannot indicate fault types. To conduct bearing fault diagnosis via statistical features, artificial neural networks and support vector machines have been applied to identify bearing faults by extracting and training the features of vibration data [22]. However, a large number of samples are required to train the fault diagnosis system with this approach.

Frequency analysis is an important technique for bearing fault diagnosis. A local defect of a bearing induces an impulse excitation each time the defect strikes other contacting parts and this results in an impulse response [17]. The impulse responses present in the signal occur at a certain frequency called the Fault Characteristic Frequency (FCF) which is proportional to the rotational frequency [9]. Therefore, bearing faults can be detected in the frequency domain. However, the collected signal is often contaminated by interferences and random noise, which smear the FCF in the frequency spectrum. Additionally, bearings often operate under time-varying speed conditions which renders the FCF also time-varying. Hence, there are two problems for bearing fault diagnosis with vibration signal processing: (1) bearing fault features/signature extraction from the contaminated signal and (2) bearing fault diagnosis under time-varying speed conditions.

For bearing fault feature/signature extraction, band-pass filters ([23]–[26]) and adaptive filters ([27]–[29]) have been widely investigated. However, band-pass based methods are sensitive to the frequencies of the signal components contained in the raw signal. The effectiveness of this method will be impaired if the frequency of the bearing fault signal and the frequency of any interference signals are similar, or if the frequencies are time-varying. Additionally, adaptive filters require a reference signal to remove the interference or noise in the signal. Therefore, for bearing fault features extraction, the challenges are: (1) developing frequency-independent and reference-free methods and (2) developing methods available for time-varying frequency cases.

For bearing fault diagnosis under time-varying speed conditions, order-tracking ([30]–[32]) is one of the most commonly used methods. The essence of order-tracking is resampling the signal according to the time-varying rotational speed so that the time-varying FCF can be converted into constant [33]. However, the accuracy of order-tracking is limited due to many factors [34]. In addition, tachometers are needed to measure the rotational speed, which may not be available due to the design of the machine or shortage of instruments. Therefore, for bearing fault diagnosis under time-varying speed conditions, the challenge is to develop resampling-free and tachometer-free methods.

1.2 Objectives

According to the challenges aforementioned, the objectives of the thesis are:

(1) Investigating the performance of the Oscillatory Behavior-based Signal Decomposition (OBSD) method, which is frequency-independent and reference-free, for bearing fault feature/signature extraction under constant speed conditions;

(2) Developing new methods based on the OBSD for bearing fault feature/signature extraction under constant speed conditions to improve the performance of the OBSD;

(3) Developing resampling-free and tachometer-free methods for bearing fault diagnosis under time-varying speed conditions based on time-frequency analysis techniques without the presence of interference signals;

(4) Developing resampling-free and tachometer-free methods for bearing fault diagnosis under time-varying speed conditions based on time-frequency analysis techniques with the presence of interference signals;

(5) Developing new methods for the bearing fault feature/signature extraction under time-varying rotational speed conditions.

1.3 Contributions of the thesis

By targeting the objectives of the thesis, the contributions of the thesis are:

(1) Development of a reliable method for bearing fault signature extraction which is frequency-independent and reference-free;

(2) Development of new methods for bearing fault diagnosis under time-varying speed conditions, which are resampling-free and tachometer-free;

(3) Development of a new method for bearing fault signature extraction under time-varying speed conditions.

1.4 Organization of the thesis

The rest of the thesis is organized as follows. In Chapter 2, methods for bearing fault diagnosis are reviewed, including methods for bearing fault signature extraction and methods for bearing fault diagnosis under time-varying speed conditions. In Chapter 3, effects of parameter selection on the OBSD for bearing fault signature extraction under constant speed condition are investigated. Then, automatic parameter selection algorithms are developed and an auto-OBSD method is proposed for bearing fault signature extraction under constant speed conditions in Chapter 4. Following in Chapter 5, a new method called Multiple Time-Frequency Curve Extraction (MTFCE) is proposed based on time-frequency analysis, and a resampling-free and tachometer-free method is proposed for bearing fault diagnosis under time-varying speed conditions without the presence of interferences. Then, another resampling free and tachometer-free method is proposed for bearing fault diagnosis under-time varying speed conditions with the presence of interferences in Chapter 6. Additionally, a new method called the short-time kurtogram is proposed for bearing fault feature extraction under time-varying speed conditions in Chapter 7. Finally, conclusions and future work are included in Chapter 8 based on all the previous chapters.

Chapter 2 Literature review

Observing the Fault Characteristic Frequency (FCF) and its harmonics in the frequency domain is one of the most effective means to detect and diagnose bearing faults [7]. The local defect of a bearing induces an impulse excitation each time the defect strikes other contacting parts and this results in an impulse response [9]. Fig. 2.1 shows the typical structure of a rolling element bearing. These impulse responses appear at a certain characteristic frequency which is proportional to the shaft rotational speed and depends on the location of the defect. This frequency is known as the FCF.

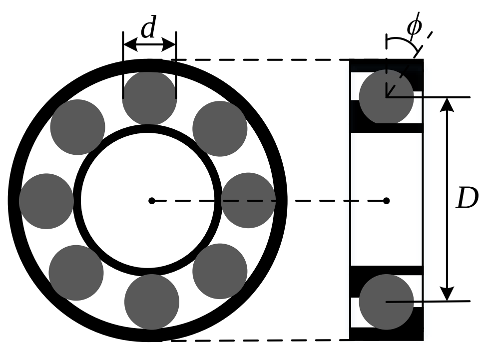

Fig. 2.1 A typical structure of rolling element bearing (cited from [9])

The FCFs for certain types of faults are given by the following equations [9]. For the ball-pass frequency, outer race

$$\text{BPFO} = \frac{nf_r}{2}\left(1 - \frac{d}{D}\cos\phi\right) \tag{2.1}$$

For the ball-pass frequency, inner race:

$$\text{BPFI} = \frac{nf_r}{2}\left(1 + \frac{d}{D}\cos\phi\right) \tag{2.2}$$

For the fundamental train (cage) frequency:

$$\text{FTF} = \frac{f_r}{2}\left(1 - \frac{d}{D}\cos\phi\right) \tag{2.3}$$

For the roller spin frequency:

$$\mathrm{RSF} = \frac{D}{2d}\left[1-\left(\frac{d}{D}\cos\phi\right)^2\right] \tag{2.4}$$

In Eq.(2.1) to (2.4), n is the number of rollers, f_r is the shaft rotational speed, d is the roller diameter, D is the pitch diameter and ϕ is the angle of the load from the radial plane.

In real applications, the collected vibration data are often contaminated by interferences and noise. Bearing fault signature extraction is often necessary prior to fault diagnosis. Additionally, bearings operate under various conditions. It could be constant speed conditions or time-varying speed conditions. Theoretically, the rationale of bearing fault diagnosis under different operational conditions should be the same, i.e., monitoring the FCF features in the frequency domain. However, the FCF becomes time-varying under time-varying conditions. Extra processing is needed to handle the time-varying FCF for bearing fault diagnosis under time-varying speed conditions. Therefore, the literature review here is organized as bearing fault signature extraction, bearing fault diagnosis under time-varying speed, and then followed by the motivation for the work in this thesis.

2.1 Bearing fault signature extraction

Bearing fault signature extraction has been widely investigated. Among existing methods, the most commonly used ones can be categorized as frequency/scale based filtering approaches, reference-based adaptive signal processing, and morphological signal decomposition. A summary of those methods, their advantages and disadvantages are given in Table 2.1.

A typical bearing fault-induced signal can be regarded as the impulse response of a one-degree-of-freedom mass-spring-damper system [35]. For a single event, the signal can be expressed as

$$s(t) = Ae^{-\beta t}\sin\left(\omega_r t\right)u\left(t\right) \tag{2.5}$$

where A is the amplitude, β is the damping coefficient, ω_r is the excited resonance frequency which is determined by the vibration system, and $u(t)$ is the unit step function. An example is shown in Fig. 2.2(a) with ω_r=2000*2π rad/s. The impulse response given in Eq.(2.5) appears periodically at the FCF. Hence, the bearing fault-induced signal can be modeled as [36]

$$x(t)=\sum_{m=1}^{M}A_m e^{-\beta\left(t-mT_p-\sum_{i=1}^{m}\tau_i\right)}\sin\left[\omega_r\left(t-mT_p-\sum_{i=1}^{m}\tau_i\right)\right]u\left(t-mT_p-\sum_{i=1}^{m}\tau_i\right) \tag{2.6}$$

where M is the total number of impulses which is determined by the signal length T and the FCF (f_c), A_m is the amplitude of the m^{th} fault impulse response, T_P is the reciprocal of f_c, τ_i is the random slippage during each T_P, usually estimated as 0.01 T_P ~0.02 T_P. The FCF can be calculated from Eq.(2.1), (2.2), (2.3), or (2.4) according to the bearing fault type. Similarly, an example of the defective bearing signal series is shown in Fig. 2.2(b) with f_c=35Hz. The fault-induced signal can be regarded as a sinusoidal signal with frequency ω_r, amplitude-modulated by an exponential decay with frequency f_c.

Table 2.1 A summary of methods for bearing fault signature extraction

No.	Categories	Methods examples	Pros	Cons
1	Band-pass filtering	Indices guided band-pass filtering, such as spectral-kurtosis [23], smoothness index [25], discrete wavelet transform [37], etc.	Straightforward, fast, reference-free	Frequency-dependent, cannot be directly used for time-varying speed cases
2	Reference-based filtering	Adaptive noise cancellation [38], self-adaptive noise cancellation [39], etc.	Frequency independent, can be used for time-varying speed case; adaptive	Relies on the reference signal (extra instruments are needed)
3	Morphological filtering	Oscillatory Behavior-based Signal processing (OBSD) [40]	Frequency independent, reference-free, potentially effective for time-varying speed cases	Method-related parameters have to be tuned

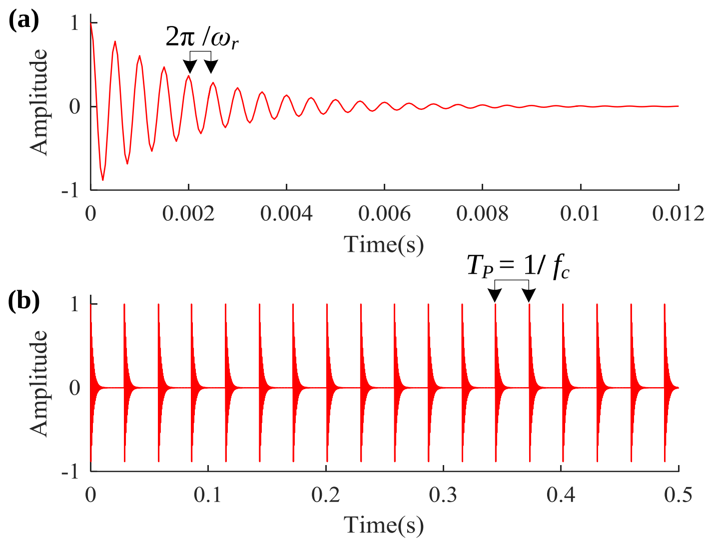

Fig. 2.2 Bearing fault-induced signal

The High-Frequency Resonance Technique (HFRT) is one of the most widely used methods to detect and diagnose bearing faults [7]. As shown in Fig. 2.3(b), the excited resonance frequency ω_r and sidebands ($\omega_r \pm n(2\pi f_c)$ rad/s where n is an integer) show clear peaks in the frequency spectrum of the bearing fault-induced signal (shown in Fig. 2.3(a)). The fault can be diagnosed based on this characteristic. However, a high resolution is required to analyze the frequency range.

To shift the frequency analysis from a high-frequency range to a relatively low-frequency range, envelope analysis, also known as amplitude demodulation, was investigated [9]. The Hilbert transform is one of the most frequently used amplitude demodulation methods for bearing fault diagnosis [41]. For a signal $x(t)$, the Hilbert transform is defined as [42]

$$H\left[x(t)\right]=\frac{1}{\pi}\int_{-\infty}^{+\infty}x(t)\frac{1}{t-\tau}d\tau \tag{2.7}$$

By demodulating the signal, the carrier frequency ω_r can be demodulated. As shown in Fig. 2.3(c), the envelope spectrum is dominated by the FCF and its harmonics. The carrier frequency ω_r and the sidebands have disappeared in Fig. 2.3(c). However, the signal is often contaminated by noise. It can be seen in Fig. 2.3(e) that the ω_r and the sidebands cannot be distinguished from the frequency spectrum when the signal contains noise (Fig. 2.3(d)). Additionally, it is difficult to recognize the FCF and its harmonics in the envelope spectrum (Fig. 2.3(f)). Therefore, pre-processing is necessary for reliable bearing fault diagnosis.

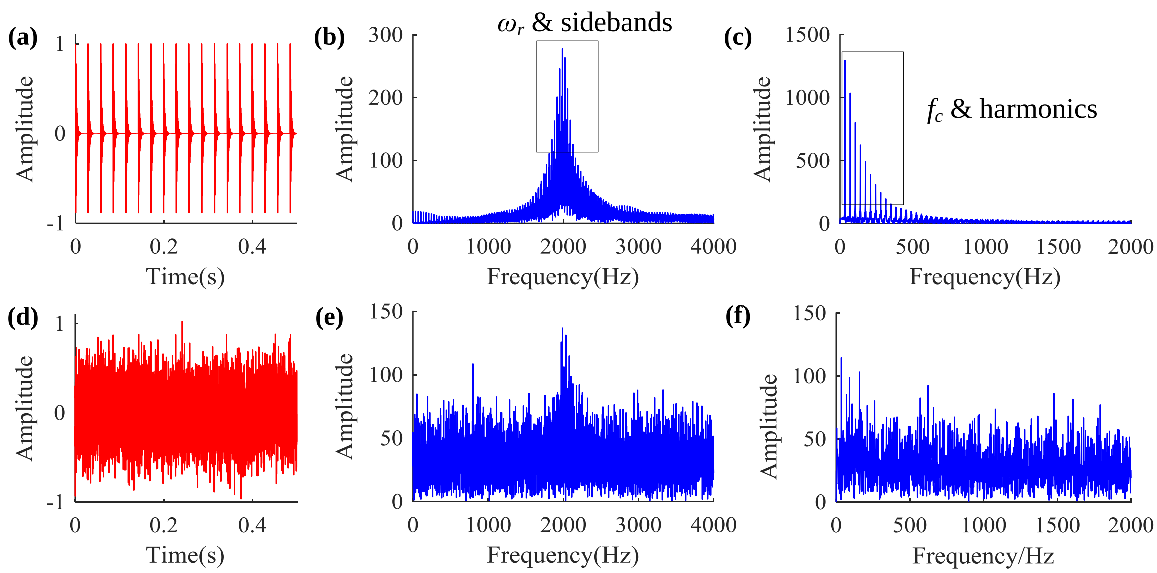

Fig. 2.3 (a) Bearing fault-induced signal, (b) its frequency spectrum and (c) envelope spectrum; (d) bearing fault-induced signal with noise, (e) its frequency spectrum and (f) envelope spectrum.

2.1.1 Frequency/scale based filtering

Band-pass filtering is one of the most commonly used techniques to enhance the bearing fault features from the blurred signal. By designing a band-pass filter with a proper center frequency and proper bandwidth, much of the noise can be effectively removed. According to the features of the bearing fault-induced signal, the center frequency should be close to the resonance frequency. Hence, plenty of existing research has been devoted to the selection of the optimal band for the band-pass filters [43].

Optimal band-pass filters for bearing fault detection are often designed based on certain criteria such as kurtosis and Spectral Kurtosis (SK) ([23], [44]–[46]). Kurtosis is the fourth standardized moment, defined as

$$kurt(x(t)) = \frac{E\left\{\left(x(t) - \mu\right)^4\right\}}{\sigma^4} - 3 \tag{2.8}$$

where μ and σ are the mean and standard deviation of signal $x(t)$, respectively, and $E\{\cdot\}$ is the expectation operator. Kurtosis can be used to quantify the impulsiveness or 'peakiness' of a signal. Since bearing fault-induced pulses are peaky, the maximum value of kurtosis can guide the selection of the center frequency and the bandwidth of the optimal filter. The SK is defined as the kurtosis of the frequency spectrum of a signal series [46], which can be employed to measure the peakiness of the signal in the frequency domain. Kurtosis and the SK are widely used to design optimal band-pass filters for bearing fault diagnosis since the ideal bearing fault-induced signal is spiky both in time and frequency.

The Discrete Wavelet Transform (DWT) is also widely used for signal de-noising because of its multi-resolution properties ([37], [47]–[49]). Similar to a frequency-based band-pass filter, the DWT can be used to decompose signals into different scales. The SK is also used to design optimal filters for bearing fault detection via the DWT [50]. In addition, a smoothness index has been employed to guide the selection of the parameters of the wavelet transform for bearing signal de-noising [26].

It is worth mentioning that an energy operator has been proposed for bearing signal demodulation, which has been shown to be usefulfor signals contaminated by interferences due to an energy shift at a high resonance frequency [51]. Additionally, methods based on the energy operator have been developed to handle higher degree interferences and noise ([52]–[54]). However, the energy operator is also frequency dependent when it is used to remove interference. The effectiveness of the interference cancellation is yet to be investigated if the frequency of the interference is also high.

From the above, frequency/scale based filtering has been commonly used for bearing signal enhancement due to its simplicity. However, the effectiveness of the frequency/scale based method is questionable if the frequency of the bearing fault-induced signal and the frequency of the interference are close to each other.

2.1.2 Reference-based noise cancellation

In addition to random background noise, the bearing signal is often contaminated by discrete frequency interferences, such as any signal transmitted from other vibration sources of the equipment. Such an interference can be modeled as [29]

$$g(t)=\sum_{i=1}^{G} B_i \sin\left(2\pi i f_g\right) \tag{2.9}$$

where G is the total number of harmonics of the interference, B_i is the amplitude of the i^{th} harmonic, and f_g is the frequency of the interference. As shown in Fig. 2.4(b), the frequency spectrum of the bearing signal contaminated by an interference (Fig. 2.4(a)) is dominated by the frequency of the interference and its harmonics. Additionally, the envelope spectrum, shown in Fig. 2.4(c), is also dominated by the frequency of the interference and its harmonics. Thus, methods for interference cancellation have also been investigated.

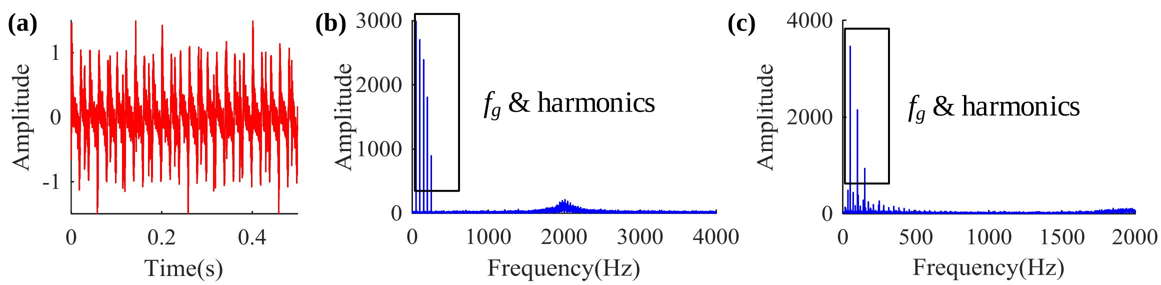

Fig. 2.4 (a) Bearing fault-induced signal with interference, (b) its frequency spectrum and (c) envelope spectrum

Adaptive filtering is widely used to deal with the cancellation of interferences for bearing fault diagnosis [38]. The idea of the adaptive filtering is to find the optimal linear filter to estimate the desired signal with the assistance of some auxiliary information. A typical structure of Adaptive Noise Cancellation (ANC) used for bearing fault detection is shown in Fig. 2.5 [27]. The observed signal $d(n)$ is the bearing signal $x(n)$ contaminated with interference $g(n)$. The filtering process is essentially one of estimating the desired signal by using a series of past data. The interference is estimated by a transversal filter with a reference signal which is correlated to the interference signal as the input. The bearing signal is then estimated as $\hat{x}(n)$ by subtracting the output of the filter from the observed signal, which is also called the error of the adaptive filter $e(n)$. The filter is adaptively tuned according to the error. The optimal filter can be determined by minimizing the error between the signal $x(n)$ and the estimated signal $\hat{x}(n)$ if the original signal $x(n)$ is uncorrelated to the interference $g(n)$ and the reference $u(n)$. To accommodate a non-stationary environment, the adaptive algorithm is used to adjust the coefficients of the filter adaptively. The most commonly applied algorithms are Least-Mean-Square (LMS) and Recursive Least-Square (RLS) algorithms, which make the filter converge to the Wiener filter recursively [55]. ANC can be used to remove the interference from a contaminated signal for bearing fault detection [28]. Additionally, ANC can also be employed for non-stationary circumstances, such as bearing fault detection under time-varying rotational speed [29]. However, the reference signal related to the interference is not always measurable due to the unavailability of sensors.

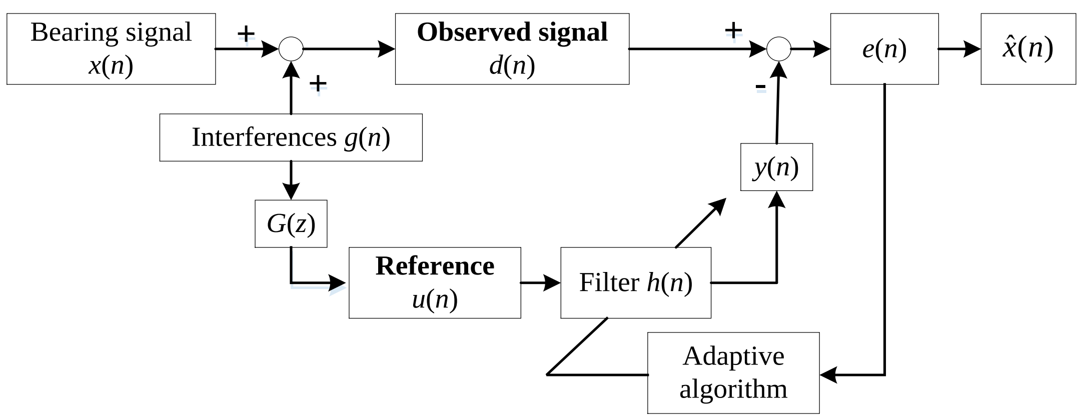

Fig. 2.5 Schematic diagram of adaptive noise cancellation

To solve the problem of the reference signal, the Self-Adaptive Noise Cancellation (SANC), also known as Adaptive Line Enhancement (ALE), was developed based on the ANC ([55], [56]). The schematic diagram of the SANC is shown in Fig. 2.6. The reference signal is replaced by a delayed version of the observed signal [9]. If the $x(n)$ is uncorrelated to $g(n)$ and the delayed signal $d(n-k)$, the solution strategy of the SANC is the same as with the ANC. The SANC based on the LMS algorithm can be implemented to remove the interference from an observed signal ([57], [58]). To reduce the computational cost of the algorithms in the time domain, a frequency domain based algorithm was developed [59]. However, the SANC is effective for interference cancellation only if the delayed length k is properly selected, so that $x(n)$ is uncorrelated to $x(n-k)$.

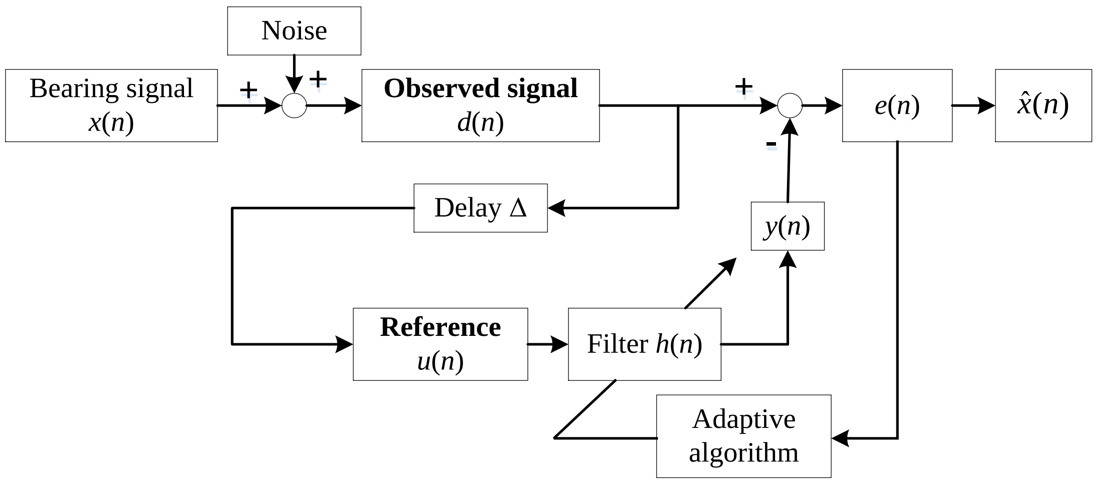

Fig. 2.6 Schematic diagram of self-adaptive noise cancellation

Autoregressive (AR) spectral estimation is a method that is also based on the theory of adaptive filtering. Given a sample of a zero-mean Gaussian stationary time series, its spectral density can be estimated by an autoregressive spectral density estimator [60]. By recursively solving the Yule-Walker equations, the estimated AR coefficients can be determined. Combining the discrete wavelet packet analysis, the AR spectrum of the envelope signal is employed to exclude contaminating sources of vibration for bearing fault diagnosis [61]. In addition, an AR signal model is obtained by training the healthy bearing signals to track changes during the development of bearing faults.

From the above, it is clear that reference-based noise cancellation methods are strongly dependent on the availability of a reference signal. The reference signal is usually collected with an extra sensor, which increases the cost. Alternatively, the reference signal can be constructed from prior knowledge, however, this is not always available.

2.1.3 Morphological filtering

More recently, a resonance-based signal decomposition method, Oscillatory Behavior-based Signal Decomposition (OBSD), was proposed by Selesnick [40]. It is a potential technique for bearing fault signature extraction under time-varying rotating speed and time-varying load conditions. Unlike band-pass based methods, with this method, the signal is decomposed according to its oscillatory behavior, as defined by a Q-factor. Thus, it is independent of frequency features. The Tunable Q-factor Wavelet Transform (TQWT) was developed to provide an overcomplete basis where basis terms have the same oscillatory behavior ([62], [63]). Since the bearing fault-induced signal is composed of impulsive responses which can be regarded as a low-oscillatory component with low Q-factor that can be easily separated from high oscillatory components, the application of the TQWT to bearing fault signal extraction has been explored recently ([64]–[66]). However, it still focuses on locating the optimal sub-band of the TQWT for bearing fault-induced signals. To take advantage of resonance-based signal decomposition, the application of Morphological Component Analysis (MCA) based on the TQWT, which is the essence of the OBSD, has recently been further investigated for bearing fault signature extraction. The bearing fault signature is regarded as an impulsive low-oscillatory component whereas the interference is usually a smoother signal which can be considered as a high-oscillatory component. Thus, the bearing signal is decomposed into a low-oscillatory component (bearing transient signal) and a high-oscillatory component (interferences) by implementing the OBSD [67]. To reduce the influence of noise on the MCA, the Ensemble Empirical Mode Decomposition (EEMD) was jointly used with the OBSD [68]. Moreover, an iterative OBSD was developed to extract the bearing fault signature from a signal obscured by a series of interferences [69]. All these applications have shown that the OBSD with well-selected parameters can be used to effectively extract the bearing fault feature from the raw vibration signal contaminated by interferences and noise. However, all the investigations were conducted with data collected under constant operating conditions.

From the above, it can be seen that the application of the OBSD on bearing fault signature extraction has not been thoroughly investigated, especially under time-varying cases. Furthermore, there are 10 parameters involved in the OBSD that must be properly selected. Poorly-selected parameters inevitably undermine the usefulness of the OBSD output and mislead fault detection decisions. A systematic OBSD parameter selection has not been adequately investigated in the literature. A subjective or trial-and-error approach is obviously not acceptable for quick and reliable applications.

2.2 Bearing fault diagnosis under time-varying speed

Time-varying rotational speed conditions make the vibration signals of rotating machines non-stationary. This introduces additional difficulty to bearing fault signature extraction. Additionally, the FCF is no longer constant since it is proportional to the shaft rotational frequency. In such cases, bearing faults cannot be detected or diagnosed by a frequency spectrum obtained via the Fourier Transform. Thus, a time-varying FCF problem, including the bearing fault signature extraction step, must be solved for bearing fault diagnosis.

There are two approaches to non-stationary vibration signal processing for bearing fault diagnosis. One is to use order tracking resampling to convert the non-stationary signal collected at a constant time increment (Δt) into a stationary signal at constant angular increment ($\Delta\theta$) [33]. The other approach is to use time-frequency analysis to reveal the local features in both time and frequency domains simultaneously [70]. A summary of methods for bearing fault diagnosis under time-varying speed conditions is given in Table 2.2.

Table 2.2 A summary of methods for bearing fault diagnosis under time-varying speed conditions

No.	Categories	Methods examples	Pros	Cons
1	Order tracking (resampling)	computed order tracking [34]	Make methods used for constant speed case still useful	Extra instruments are needed to measure the time-varying speed, accuracy is limited by interpolation methods.
2	Time-frequency analysis	Short-Time Fourier Transform (STFT) [71], wavelet transform [72], etc.	Instantaneous FCF (IFCF) can be directly shown in the Time-Frequency Representation (TFR), resampling-free	Time-frequency curves need to be processed

2.2.1 Order tracking resampling

The essence of order tracking is the resampling of the original vibration signal at a constant angle increment which converts the non-stationary signal in the time domain into a stationary signal in the angular or order domain. Using the ratio of FCF to shaft rotational frequency, bearing faults can be diagnosed via the envelope spectrum of the resampled signal following the same method used with constant speed cases.

There are two types of order tracking according to the devices and techniques used to resample the signal: hardware-based order tracking and computed order tracking [34]. The former directly resamples the vibration signal at constant angular increments, i.e. proportional to the rotating speed, using analog instrumentation. Normally, a ratio synthesizer is employed to generate a signal proportional to the rotational speed measured by a tachometer. Then, the analog vibration signal is resampled at the sampling rate controlled by the ratio synthesizer. The hardware method necessitates high equipment cost and may not be able to follow rapidly changing speeds. Hence, the computed order tracking, a fully digital technique, was developed to numerically obtain the resampling rate from the tachometer. Assuming the shaft rotates at constant angular acceleration, the shaft angle can be described by

$$\theta(t) = b_0 + b_1 t + b_2 t^2 \tag{2.10}$$

The coefficients b_0, b_1, and b_2 can be determined by the tachometer by measuring at three successive times t_0, t_1, and t_2 for the angles 0, $\Delta\Phi$, and $2\Delta\Phi$. If the signal needs to be resampled at $\Delta\theta$, then the resample times within the time interval t_0 to t_2 can be determined as

$$t = \frac{1}{2b_2}\left[\sqrt{4b_2\left(k\Delta\theta - b_0\right) + b_1^2} - b_1\right], \qquad k=1,2,3... \tag{2.11}$$

The signal is then successively resampled with the update of the new measured time for $\Delta\Phi$.

The computed order tracking for bearing fault diagnosis under time-varying rotational speed has been investigated by many researchers ([30]–[32], [72]). An intelligent order tracking system was developed by applying the RLS algorithm and the Kalman filter to extract the order amplitudes of the vibration signal [30]. In addition, the capability of order tracking for discrete-random separation has been investigated for fault diagnosis of rotating machinery under variable speed conditions [31]. Moreover, a new procedure was developed by combining envelope analysis with computed order tracking for bearing fault diagnosis under variable operating conditions [73]. However, computed order tracking is dependent on the acquisition of the instantaneous rotational speed of the shaft. To make the order tracking available without tachometers, a Short-Time Fourier Transform (STFT) based instantaneous rotating frequency estimation method has been carried out for bearing fault diagnosis under time-varying rotational speed conditions [32]. However, the overall accuracy of the computed order tracking method is affected by factors such as digital filtering, the chosen interpolation method, noise, and data block size [34]. For example, by resampling the signal at different frequencies, the number of samples for a segment has to increase or decrease, which may change the character of the original signal.

2.2.2 Time-frequency analysis

Time-frequency analysis is based on Time-Frequency Representation (TFR) of a signal in a time-frequency-amplitude/energy density 3D space. Similar to the constant speed case, the variable FCF and its harmonics exhibit peaks in the TFR which can be used for bearing fault diagnosis under time-varying speed conditions.

Various approaches to TFR have been developed. Linear TFR is commonly used for machinery fault diagnosis under time-varying operational conditions, such as the Short-Time Fourier Transform (STFT) [74] and Wavelet Transform (WT) [72]. The idea of linear TFR is to use orthogonal basis functions to estimate the intervals of the signal. Hence, these methods are free from cross-terms and can be implemented quickly. However, the time and frequency resolution are limited by the Heisenberg uncertainty principle. To address the resolution issue, the quadratic/ bilinear time-frequency distribution was proposed and applied to machinery fault detection such as with the Wigner-Ville distribution [75]. However, these methods suffer from the cross-term interference if the signals are of multi-component. Many improved methods such as Cohen distribution [76], Affine class distribution [77], and Reassigned distribution [78] were developed to suppress the cross-terms and can be used for machine fault detection. However, this comes at the expense of reducing the time-frequency resolution. Among these methods, linear TFR is the most frequently employed for bearing fault diagnosis. Thus, linear TFR is further reviewed in the rest of this sub-section.

Linear TFR can be regarded as a projection of a signal on a defined basis in the time-frequency domain. All linear TFRs satisfy the superposition principle, which implies that the TFR of a signal's combined components is equal to the sum of TFRs of each component. This property eliminates the cross-term problem compared to quadratic TFR. The STFT and the WT are two basic linear TFRs. They have been widely applied to fault diagnosis of rotating machinery, including bearings and gears. Although high resolutions cannot be obtained simultaneously in the time and frequency domain, the linear TFR is still of great importance.

The Short-time Fourier Transform (STFT) is the earliest time-frequency method, applied by Gabor [79] in 1946 to speech communication. The Short-time Fourier transform is developed simply by adding a window function to the Fourier transform. The mathematical expression is given by [71]

$$STFT_x(\tau, f) = \int_{-\infty}^{+\infty} x(t)\, w(t-\tau)\, e^{-j\omega t}\, dt \tag{2.12}$$

where $w()$ is the window function. The time resolution and the frequency resolution are determined by the width of the window. Normally, a wider window provides good frequency resolution, but poor time resolution and vice versa. Due to the simplicity of the STFT, it is widely used for bearing fault diagnosis under variable speed conditions in combination with other methods. For example, the synchrosqueezing transform is applied to visually improve the resolution of the TFR obtained via STFT for bearing fault diagnosis ([80], [81]). Additionally, a generalized demodulation is employed to convert the time-frequency component into a linear path parallel to the time axis for bearing fault diagnosis [81]. However, since the duration of the basis of STFT is for all time and the window width is constant, the time-frequency resolution is uniform throughout the time-frequency plane. Thus, the application of STFT is limited by its resolution.

The Wavelet transform (WT) is a multi-resolution transform that employs a wavelet and scaling function as the basis. It gives good time resolution at high frequencies and good frequency resolution at low frequencies. The continuous WT is defined as [82]

$$WT_x(s,\tau)=\frac{1}{\sqrt{s}}\int_{-\infty}^{+\infty}x(t)\Psi_{s\tau}\left(\frac{t-\tau}{s}\right)dt \tag{2.13}$$

where *s* is the scale related to the frequency defined as the reciprocal of frequency, τ is the time shift related to the location of the window and corresponds to time information in the transform domain, and $\Psi()$ is the wavelet basis function. The length of window function of the WT varies by translating and dilating the wavelet function, whereas the width of the window of the STFT is fixed. Thus, the WT provides better time-frequency resolution compared to the STFT. However, it shows poor frequency resolution at high frequencies and poor time resolution at low frequencies. To improve the time-frequency resolution, methods have been proposed based on the WT, such as wavelet package transform which decomposes the signal not only on the low-pass side but both sides [83]. In addition, the performance of the WT also depends on the wavelet function. The problem of how to effectively choose a wavelet function is still to be addressed.

Compared with the order tracking resampling, utilizing TFR for bearing fault diagnosis under time-varying speed conditions is more straightforward since the signal does not need to be resampled. However, bearing fault signature extraction is necessary for bearing fault diagnosis with either the order tracking resampling or the time-frequency analysis as long as the signal is contaminated by interference and noise.

2.3 Motivation

According to the literature, there are two main problems for bearing fault diagnosis: (1) one is the bearing fault signature extraction under constant speed and time-varying speed conditions; (2) the other one is bearing fault diagnosis or bearing fault recognition under time-varying speed conditions.

For the bearing fault signature extraction, the OBSD method is frequency-independent and reference-free, which is superior to the band-pass filtering and adaptive filtering. However, 10 method-related parameters have to be tuned when the OBSD method is applied to bearing fault signature extraction. Thus, the parameter selection for the OBSD is worth investigation. Additionally, for bearing fault signature extraction under time-varying speed conditions, the OBSD method is potentially effective since it is frequency-independent. Moreover, it is also possible to improve the availability of band-pass filters for bearing fault signature extraction under time-varying speed conditions.

For bearing fault diagnosis or bearing fault recognition under time-varying speed conditions, methods based on the time-frequency analysis are worth investigating since resampling is not necessary when the signal is analyzed in the time-frequency domain. Additionally, the tachometer may not be needed if the time-varying speed can be extracted from the result processed in the time-frequency domain.

Chapter 3 Effects of parameter selection on OBSD for bearing fault signature extraction

This chapter addresses objective 1, i.e. the effects of the parameter selection on the performance of the OBSD was investigated for the bearing fault signature extraction under constant speed conditions.

The content of this chapter has been published in the *Proceedings of the ASME Design Engineering Technical Conference*.

Huan Huang, Natalie Baddour, and Ming Liang. "Effects of Parameter Selection on Oscillatory Behavior-Based Signal Decomposition for Bearing Fault Signature Extraction" *ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers, 2016. Paper No. DETC2016-59391, pp. V008T10A038; 7 pages doi:10.1115/DETC2016-59391

3.1 Abstract

Oscillatory Behavior-based Signal Decomposition (OBSD) is a new technique which employs Morphological Component Analysis (MCA) and the Tunable Q-factor Wavelet Transform (TQWT) to decompose a signal into components consisting of different oscillatory behaviors rather than different frequency bands or scales. Due to the low oscillatory transients of bearing fault-induced signals, this method shows promise for applications related to effectively extracting bearing fault signatures from raw signals contaminated by interferences and noise. In this chapter, the application of OBSD to bearing fault signature extraction is investigated. It is shown that the quality of the results obtained via the OBSD is highly dependent on the selection of method-related parameters. The effects of each parameter on the performance of the OBSD for bearing fault signature extraction are investigated. The analysis is also validated by implementing the OBSD on experimental data collected from a test rig with a defective bearing.

3.2 Introduction

Bearings are one of the most commonly used components associated with mechanical machines. Failure in bearings degrades the performance of the machine or even results in a failure of the machine. Thus, bearing fault detection is critical to ensure good operating conditions for machines. Vibration signal analysis is a widespread means to detect bearing faults, since for a given fault type such as inner race fault or outer race fault, the impulses induced by a bearing fault appear at a specific frequency, referred to as the fault characteristic frequency (FCF) [9].

However, in practice the measured bearing vibration signal is often contaminated by interference transmitted from other sources such as gears and background noise. The FCF and its harmonics are then buried in the frequency spectrum and therefore difficult to detect when the interferences are strong. Therefore, bearing fault signature extraction is an important pre-processing step for bearing fault detection.

Many approaches have been proposed for bearing fault feature enhancement and these can be roughly classified into reference-based signal enhancement and bandpass filter-based signal enhancement. The reference-based methods employ adaptive signal processing techniques, which are able to remove or retain the component correlated to the reference ([57], [28]). These methods are free from frequency information. However, reference signals are required, which are not always available, especially in real-world applications. Bandpass filter-based signal enhancement is widely applied to extract a signal from given frequency bands. An example of this is the popular Spectral Kurtosis (SK)-based filtering ([24], [84]). However, the optimal band is easily influenced by variation in the frequency of the interference and thus the optimal band cannot always be reliably found.

Oscillatory Behavior-based Signal Decomposition (OBSD) is a new technique developed to decompose a signal into components consisting of a low oscillatory level component and a high oscillatory level component ([40], [62]). The method is frequency-independent and reference-free. Hence, it avoids the disadvantages of the methods mentioned above. Impulses generated by a bearing fault are impulsive, which can be considered as a low oscillatory component. However, the interference is usually composed of smooth oscillations which can be regarded as a high oscillatory component ([69], [85]). Hence, it is hypothesized that the OBSD can be applied to extract bearing fault signatures from a signal smeared by interference and noise.

However, OBSD is not a simple method of signal decomposition and its use involves the selection of several parameters. In this chapter, the effect of parameter selection on the effectiveness of the OBSD for bearing fault feature extraction is investigated.

3.3 Oscillatory Behavior-based Signal Decomposition (OBSD)

The OBSD is developed based on the idea of Morphological Component Analysis (MCA) that optimally separates a signal into components consisting of different morphological features, i.e. different oscillatory behaviors. Additionally, the Tunable Q-factor Wavelet Transform (TQWT) is employed to supply wavelet bases with adjustable oscillatory behavior. The steps of the OBSD are: (1) first, separately generate two sets of wavelets, one with a low oscillatory level and the other with a high oscillatory level; (2) apply the MCA to obtain the optimal wavelet coefficients for the low oscillatory component and the high oscillatory component, respectively; (3) finally, obtain the low oscillatory component and the high oscillatory component via the inverse TQWT with the optimal wavelet coefficients found in the previous step.

The TQWT is carried out based on the fundamental premise that the Q-factors of a waveform at different scales are the same. The Q-factor of a waveform is defined as the ratio of its center frequency to its band-width [62], which is a measure of the oscillatory level of the waveform. Thus, the TQWT is adopted to generate the wavelet bases of the specific oscillatory behavior, which is the foundation for the OBSD. The bases are dependent on the selection of the associated parameters Q-factor (Q), redundancy (r) and number of levels (P). An example is shown in Fig. 3.1 where Q=3, r=3 and P=15. Fig. 3.1(a) shows the wavelets with Q-factor Q=3 at 15 different scales, and the redundancy of the scales is defined as r=3. Additionally, Fig. 3.1(b) demonstrates the corresponding 15 frequency spectra of the wavelets. Since the signal is decomposed into two components, two sets of wavelet bases are needed to support the OBSD, one with a low Q-factor and the other with a high Q-factor. Thus, six parameters are involved in the generation of the wavelets for the OBSD, Q_l, r_l and P_l for the low Q-factor bases, and Q_h, r_h and P_h for the high Q-factor bases, respectively. With these wavelets, the TQWT is applied to obtain the wavelet coefficients of a signal. Similarly, the signal can be reconstructed via the inverse TQWT from the wavelet coefficients. Details about the implementation of the TQWT and the inverse TQWT can be found in Appendix A.

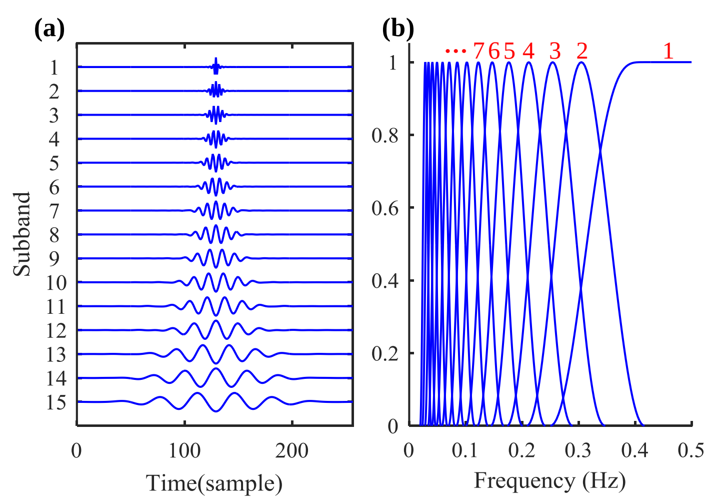

Fig. 3.1 Wavelet bases and their frequency spectra with Q=3, r=3 and P=15

The MCA [40] is essentially an optimization problem to obtain the optimal expressions of the desired components. For the OBSD, the objective function is given by Eq.(3.1) and is composed of the regularized l_1–norm of the wavelet coefficients of the low and high Q components, and the l_2-norm of the residual. The optimal wavelet coefficients are found by minimizing the given objective function.

$$J\left(\mathbf{w}_l, \mathbf{w}_h\right) = \lambda_l \left\|\mathbf{w}_l\right\|_1 + \lambda_h \left\|\mathbf{w}_h\right\|_1 + \left\|\mathbf{x} - \mathbf{S}_l \mathbf{w}_l - \mathbf{S}_h \mathbf{w}_h\right\|_2^2 \tag{3.1}$$

where $J(\mathbf{w}_l.\mathbf{w}_h)$ is the objective function to be minimized, $\mathbf{w}_l$ are the wavelet coefficients of the low Q-factor component, $\mathbf{w}_h$ are the wavelet coefficients of the high Q-factor component, $\mathbf{x}$ is the original signal, $\mathbf{S}_l$ are the wavelet bases of the low Q-component, $\mathbf{S}_h$ are the wavelet bases of the high Q-factor component, λ_l is the regularization parameter of the low Q-factor wavelet coefficients, and λ_h is the regularization parameter of the high Q-factor wavelet coefficients. The minimization of the l_1–norm terms makes the oscillatory behaviors of the low-Q and high-Q components close to the oscillatory behaviors of the corresponding wavelet bases as much as possible. At the same time, the minimization of the l_2–norm term keeps the residual as small as possible. To solve the optimization problem, the Split Augmented Lagrangian Shrinkage Algorithm (SALSA) is applied. The obtained iterative solution is given by Eq.(3.2), (3.3), (3.4), (3.5) and (3.6). The detailed derivation of the solution can be found in [40].

$$\mathbf{u}_i^{k+1} = soft\left(\mathbf{w}_i^k + \mathbf{d}_i^k, \frac{\lambda_i}{2\mu}\right), \quad i = l, h \tag{3.2}$$

$$\mathbf{b}_i^{k+1} = \mathbf{u}_i^{k+1} - \mathbf{d}_i^k, \quad i = l, h \tag{3.3}$$

$$\mathbf{c}^{k+1} = \frac{1}{\mu + 2}\left(\mathbf{x} - \mathbf{S}_h \mathbf{b}_h^{k+1} - \mathbf{S}_l \mathbf{b}_l^{k+1}\right) \tag{3.4}$$

$$\mathbf{d}_i^{k+1} = \mathbf{S}_i^T \mathbf{c}^{k+1}, \quad i = l, h \tag{3.5}$$

$$\mathbf{w}_i^{k+1} = \mathbf{d}_i^{k+1} + \mathbf{b}_i^{k+1}, \quad i = l, h \tag{3.6}$$

In Eq.(3.2)-(3.6), k is the number of iterations, μ is the penalty factor, $\mathbf{w}_l^k$ and $\mathbf{w}_h^k$ are the wavelet coefficients of low-Q and high-Q by the k^{th} iteration, $\mathbf{u}_i$, $\mathbf{d}_i$, $\mathbf{b}_i$ and $\mathbf{c}_i$ are all auxiliary variables for the iterative algorithm. As can be seen, the result is dependent on the penalty factor μ and the total number of iterations K. The optimal wavelet coefficients of low-Q and high-Q are taken as the result given in Eq.(3.6) after K iterations.

Once the optimal wavelet coefficients are determined by the MCA, the signal decomposition is obtained via inverse TQWT given as

$$\hat{\mathbf{x}}_i = TQWT^{-1}(\mathbf{w}_i^K), \quad i = l, h \tag{3.7}$$

Considering the related parameters, the full result is then given by

$$\hat{\mathbf{x}}_i = OBSD(\mathbf{x}, Q_l, r_l, P_l, Q_h, r_h, P_h, \lambda_l, \lambda_h, \mu, K) \;\; i = l, h \tag{3.8}$$

3.4 Application of the OBSD for bearing fault signature extraction

A bearing fault-induced signal includes impulses appearing at the FCF, which can be considered as low oscillatory behavior. However, the interference such as that arising from the gear meshing signal is composed of continuous oscillations, which can be regarded as high oscillatory behavior. Thus, the OBSD can be applied to separate the bearing fault-induced signal from the original signal mixed with interference.

A bearing fault-induced signal is usually modeled as [36]

$$x(t) = \sum_{m=1}^{M} A_m e^{-\beta\left(t - mT_p - \sum_{i=1}^{m}\tau_i\right)} \sin\left[\omega_r\left(t - mT_p - \sum_{i=1}^{m}\tau_i\right)\right] u\left(t - mT_p - \sum_{i=1}^{m}\tau_i\right) \tag{3.9}$$

where M is the total number of impulses which is determined by the signal length T and the fault characteristic frequency fcf, A_m is the amplitude of the m^{th} fault impulse, $u(t)$ is the unit step function, T_P is the period of the impulses (reciprocal of fcf), β is the damping coefficient, ω_r is the excited resonance frequency, and τ_i is the random slippage during each T_P -usually considered as $0.01T_P$~$0.02T_P$. The interference is usually modeled as

$$h(t) = A_h \sum_{j=1}^{G} B_j \sin\left(2\pi j f_h\right) \tag{3.10}$$

where A_h is the amplitude of the harmonics, G is the total number of harmonics, B_j is the amplitude ratio coefficient of the j^{th} harmonic, and f_h is the fundamental frequency of the interference.

An example of the implementation of the OBSD on bearing fault signature extraction is illustrated in Fig. 3.2. The original signal, shown in Fig. 3.2(e), is a mixture of bearing fault impulses (Fig. 3.2(a)) and interference (Fig. 3.2(c)). The bearing fault signal is modeled by Eq.(3.9) with fcf=35Hz, A_m=1, β=800, ω_r=4000*2π, τ_i=0.01T_P and signal length T=0.6s. In addition, the interference is modeled by Eq.(3.10) with f_h=50Hz, A_h=0.5, G=3 and B_j= [2, 1, 0.5]. The bearing fault is generally detected by the Envelope Spectrum (ES) which is the spectrum of the envelope preceded by the computation of a Hilbert transform [9]. As shown in Fig. 3.2(b), the envelope spectrum of the bearing fault-induced signal is composed of peaks at fcf and its multiples, which can be used to detect the presence of a bearing fault. Similarly, the frequency of interference and its harmonics show peaks in the envelope spectrum of the interference, shown in Fig. 3.2(d). However, it can be seen from Fig. 3.2(f) that the envelope spectrum of the original signal is dominated by the frequency of the interference and its harmonics, which implies that the bearing fault cannot be detected without additional signal treatment. By applying the OBSD on the contaminated original signal with OBSD parameters Q_l=1, r_l=3, P_l=20, Q_h=6, r_h=3, P_h=70, λ_l=0.3, λ_h=0.3, μ=2 and K=258, the signal is decomposed into a low-Q component (Fig. 3.2(g)) and a high-Q component (Fig. 3.2(i)). The residual is also given in Fig. 3.2(k). The bearing fault-induced impulses, which is the fault signature, are clearly seen in the low-Q component (Fig. 3.2(g)). Furthermore, the envelope spectrum of the low Q-factor component, shown as Fig. 3.2(h), is dominated by the FCF and its harmonics. The bearing fault can thus be easily detected after processing. It can be seen from this example that the OBSD can be effectively used to extract the bearing fault signature from a raw signal contaminated by interference if the OBSD-related parameters are properly selected.

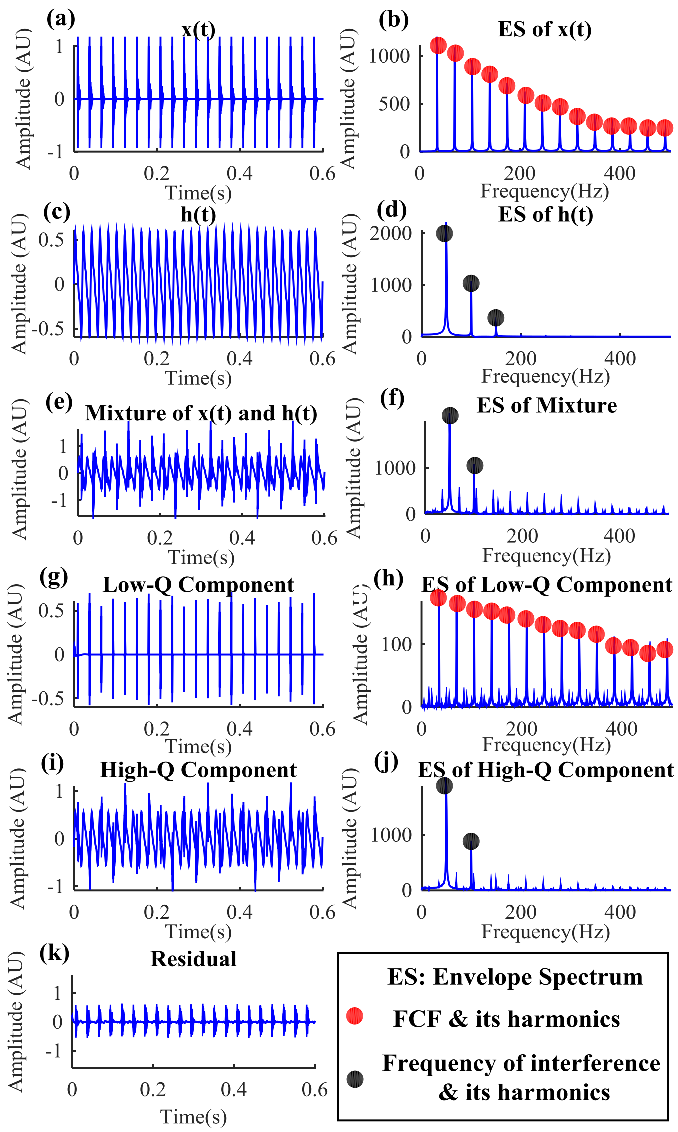

Fig. 3.2 Implementation of the OBSD on bearing fault signature extraction

3.5 Effects of parameter selection

There are ten parameters associated with the OBSD in total as given in Eq.(3.8). The effects of the parameter selection on the OBSD for the bearing fault signature extraction are discussed in this section.

The Q_l indicates the oscillatory level of the low Q-factor wavelet bases. Thus, the choice of Q_l must be matched with the oscillatory behavior of the desired signal, in this case the bearing fault-induced impulses. For applications to bearing fault detection, the Q_l should be as low as possible due to the impulsiveness of the bearing fault-induced signal. Since the Q-factor cannot be less than 1, it is reasonable to fix Q_l at 1 when applying this method for bearing signal decomposition.

Similarly, the Q_h describes the oscillatory level of the high Q-factor wavelet bases. Therefore, the Q_h should be chosen to match the oscillatory behavior of the interference. Any mismatching may lead to unreliable results. As given in [40], the correlation between the high Q-factor wavelets and the low Q-factor wavelets should be controlled to be as low as possible. Thus, the Q_h should be selected to be sufficiently greater than the Q_l. However, if Q_h is too high, the oscillatory behavior may not match the expected high oscillatory component very well.

The redundancy r is related to the extent to which the frequency responses of the transform overlap. The effects of the selection of r_l and r_h are the same. If the redundancy is increased, the density of frequency responses will be raised which means the overlapping extent of the frequency responses of the TQWT will be improved. However, the computational cost of the TQWT also climbs with an increase in redundancy [62]. Hence, growth of the redundancy boosts the overlapping extent of the frequency responses but lowers computational efficiency.

The number of levels of the wavelets P reveals the coverage of the frequency responses in the frequency domain. It should be selected to be as high as possible. Otherwise, the frequency responses may not cover the low frequencies. For example, if P is picked as 10 for the case presented in Fig. 3.2, the approximate normalized frequency range (0, 0.08) is not covered by the frequency responses. This may lead the TQWT back to a frequency-based method, instead of a fully oscillatory behavior-based method. Therefore, the P should always be chosen to be the maximum allowable value. The maximum P is given by Eq.(3.11), according to [62] as

$$P_{\max} = \left\lfloor \frac{\log\left(\dfrac{N}{4(Q+1)}\right)}{\log\left(\dfrac{(Q+1)r}{(Q+1)r-2}\right)} \right\rfloor \tag{3.11}$$

where $\lfloor A \rfloor$ is an operator that rounds the number A to the next smaller integer, N is the number of samples, Q is the Q-factor, and r is the redundancy. It can be seen that P_{max} is determined by the Q, r and N, which implies that the selection of P can be automatically completed once the Q-factor and the redundancy are confirmed.

The regularization parameters λ_l and λ_h are used to balance the residual and the sparsity of the wavelet coefficients in the objective function. With an increase of λ_l and λ_h, the objective of high sparsity of wavelet coefficients becomes stronger while the objective of small residual gets weaker. In such a case, the obtained wavelet coefficients are sparser. Then, the oscillatory behaviors of the decomposed low-Q component and high-Q component are closer to the corresponding wavelets. However, some useful information will be lost if λ_l and λ_h are increased excessively. The λ_l and λ_h also control the relative sparsity between the low-Q wavelet coefficients and high-Q wavelet coefficients. Ideally, λ_l and λ_h should be selected to be proportional to the power of the corresponding low-Q and high-Q component [40]. However, this prior knowledge is usually not available. Moreover, for bearing fault detection, only the low-Q component is of interest. Hence, it is reasonable to take λ_l=λ_h for the application of the OBSD to bearing fault signature extraction.

The penalty factor, μ, is related to convergence rate of the solution. A faster convergence rate can reduce the number of iterations (K) required for the minimization of the objective function. In other words, a faster convergence rate benefits computational efficiency. Thus, μ should be selected with the objective of the fastest convergence considering the computational cost. However, the value of μ for fastest convergence differs in different cases [86]. The convergence rate is not monotonically increasing or decreasing with the growth of μ. Therefore, the relation between μ and convergence rate can be only determined with the details of each case.

The number of iterations K should be big enough to ensure that the objective function reaches an acceptable level of minimization. Otherwise, the result may not be reliable. Theoretically, with the convergence of the solution, more iterations lead to better results. However, the computational cost is raised with an increase in K. Therefore, the selection of K is necessary to balance the convergence level of the result and computational efficiency.

3.6 Experimental evaluation

To validate the analysis of the effects of each parameter on the bearing fault signature extraction, the OBSD is applied to experimental data. The experiment is performed on a SpectraQuest machinery fault simulator (MFS-PK5M). The experimental set-up is shown in Fig. 3.3. The shaft is supported by two ER16K ball bearings, the left one is healthy and the right one is faulty with the outer race deliberately faulted [87]. The rotation of the shaft is transmitted to the gearbox by two belt sheaves. The diameter ratio of the two sheaves is 1:2.6 (the small one is connected to the shaft and the large one is connected to the gearbox). The vibration signal is collected by an accelerometer (ICP accelerometer, Model 623C01) located on the base of the test rig, and the data is sampled by a NI data acquisition board (NI USB-6212 BNC) and LABVIEW. With this configuration, the collected signal contains both the bearing fault-induced impulses and the gear meshing signal. According to the structure parameters of the bearings given in Table 3.1, the fault characteristic frequency (FCF) of the bearing outer race, i.e. Ball-pass Frequency of Outer-race (BPFO), is equal to the product of the FCF coefficient (3.57) and the shaft rotational frequency f_r [9], i.e. $fcf=3.57f_r$. In addition, the gear meshing frequency is calculated as the shaft rotational frequency divided by the diameter ratio of the sheaves and then multiplied by the number of the pinion teeth, i.e. $f_{meshing}=(18/2.6)f_r=6.92f_r$. In this case, the shaft rotational frequency was set at 15Hz, the signal sampling rate was set at 20kHz and the number of samples was 32748 (signal length was 1.6374s). Given this data, the bearing fault characteristic frequency and the gear meshing frequency are calculated as 53.55Hz and 103.8Hz, respectively. The measured signal is shown in Fig. 3.4(a) and the corresponding envelope spectrum is given in Fig. 3.4(b). It can be seen that the gear meshing frequency shows a clear peak in the envelope spectrum, while the bearing fault characteristic frequency is totally submerged. Obviously, the bearing fault cannot be detected by the envelope spectrum of the original signal.

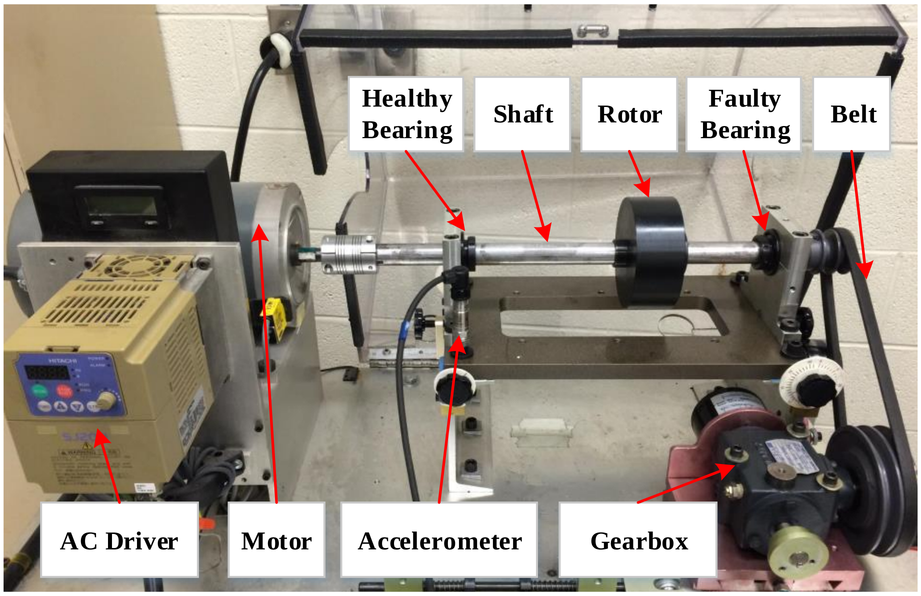

Fig. 3.3 Experimental set-up

Table 3.1 Parameters of bearings and gears

Bearing type	Pitch diameter	Ball diameter	Number of balls	BPFO
ER16K	38.52mm	7.94mm	9	$3.57f_r$
Diameter ratio of sheaves		Number of teeth	Gear meshing frequency	
1:2.6		18	$6.92f_r$	

Applying the OBSD on the collected signal, the results of the low-Q component, high-Q component and the residual are shown in Fig. 3.4(c), Fig. 3.4(e) and Fig. 3.4(g) respectively. The parameters of the OBSD are selected as Q_l=1, Q_h=8, r_l=6, r_h=6, P_l=45, P_h=180, λ_l=λ_h=0.1, μ=2 and K=200. The decomposed low-Q component is visually more impulsive than the original signal. Additionally, the envelope spectra of the low-Q and high-Q components are given in Fig. 3.4(d) and Fig. 3.4(f), respectively. It can be seen that the bearing fault characteristic frequency and its harmonics show clear peaks in Fig. 3.4(d) with no peak for the presence of gear meshing frequency. The bearing fault can be detected easily with the envelope spectrum of the low-Q component. Moreover, the envelope spectrum of the high-Q component is still dominated by the gear meshing frequency, which coincides with the objective of the OBSD. It can be concluded that the OBSD may be effectively used for bearing fault signature extraction if the parameters are properly tuned.

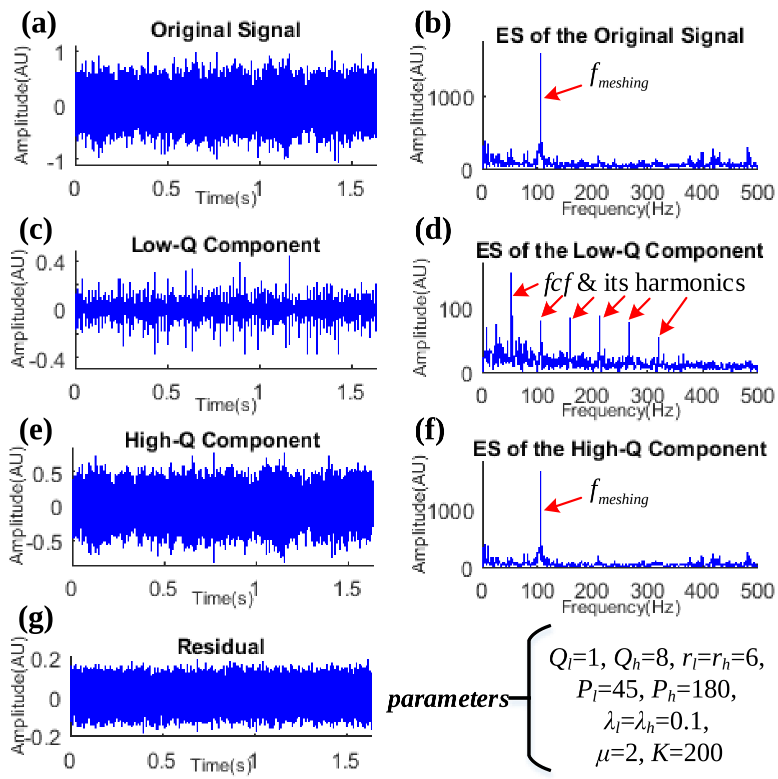

Fig. 3.4 Original signal and results of the OBSD with proper parameters

By tuning the high Q-factor to a smaller value (Q_h=2) and keeping the setting of the other parameters the same, the decomposed low-Q component and the corresponding envelope spectrum are shown in Fig. 3.5(a) and Fig. 3.5(b), respectively. Unlike the result in Fig. 3.4(d), the gear meshing frequency is still dominant in the envelope spectrum of the low-Q component, even though the bearing fault characteristic frequency shows a more distinct peak. As discussed in the previous section, the Q_h should be matched with the oscillatory level of the interference. Also, the Q_h cannot be too close to Q_l. First, the oscillatory behavior of the wavelets with the chosen Q_h=2 is too low since the gear meshing signal is continuously oscillatory. Second, Q_h=2 is too close to Q_l=1. Hence, the result in Fig. 3.5 verifies analysis in the previous section about the effects of the selection of Q_h.

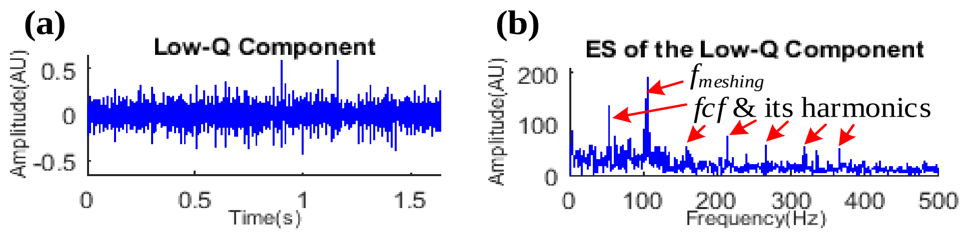

Fig. 3.5 Result of the OBSD with small high Q-factor (Q_h=2)

The result of the low-Q component and its envelope spectrum are given in Fig. 3.6 where the number of the levels of high-Q wavelets P_h was reduced from the maximum 180 to 30. It can be seen that the envelope spectrum in Fig. 3.6(b) is dominated by the gear meshing frequency, which looks the same as the envelope spectrum of the original signal in Fig. 3.4(d). This supports the earlier discussion that a small P_h may lead the OBSD to a frequency dependent method which generates unreliable results. By reducing the P_h to 30, the frequency responses of the TQWT with high-Q no longer cover the frequency of the gear meshing frequency. However, the frequency response of the TQWT with low-Q still fully cover all frequencies. Therefore, the gear meshing signal will be separated to the low-Q component via the OBSD. In such a case, the decomposed low-Q component is still contaminated by the gear meshing signal. Thus, the results shown in Fig. 3.6 confirms the previous analysis of the effect of the selection of P.

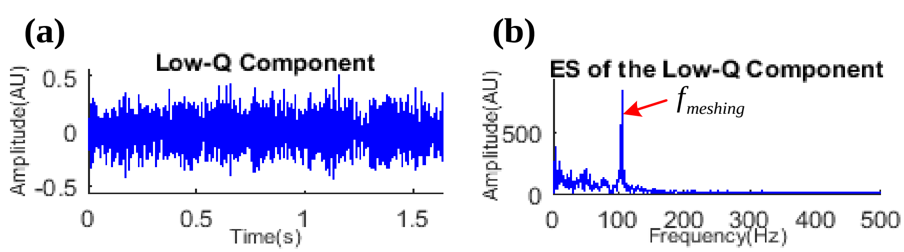

Fig. 3.6 Result of the OBSD with small number of high Q-factor wavelet levels (P_h=30)

The result of the low-Q component and its envelope spectrum are shown in Fig. 3.7(a) and Fig. 3.7(b) by decreasing λ_l and λ_h to 0.01, and in Fig. 3.7(c) and Fig. 3.7(d) by increasing λ_l and λ_h to 0.3, respectively. The result in Fig. 3.7(b) implies that the gear meshing frequency is still dominant in the envelope spectrum of the low-Q component when λ_l and λ_h are 0.01. On the other hand, from the results in Fig. 3.7(d), the gearing meshing frequency is suppressed when λ_l and λ_h are 0.3. However, the bearing fault characteristic frequency and its harmonics are not clearly shown in the envelope spectrum compared with the result in Fig. 3.4(d). It was analyzed in the previous section that greater λ_l and λ_h lead to sparser results. If λ_l and λ_h are too small, the interference cannot be fully removed in the low-Q component. However, if λ_l and λ_h are too large, the bearing fault-induced impulses will be lost in the low-Q component. In this case, $\lambda_l = \lambda_h = 0.01$ is too small to fully remove the interference from the low-Q component. Nevertheless, $\lambda_l = \lambda_h = 0.3$ is too large to keep all the bearing fault impulses in the low-Q component. Therefore, the results in Fig. 3.7 imply the effectiveness of the previous analysis about the selection of λ_l and λ_h.

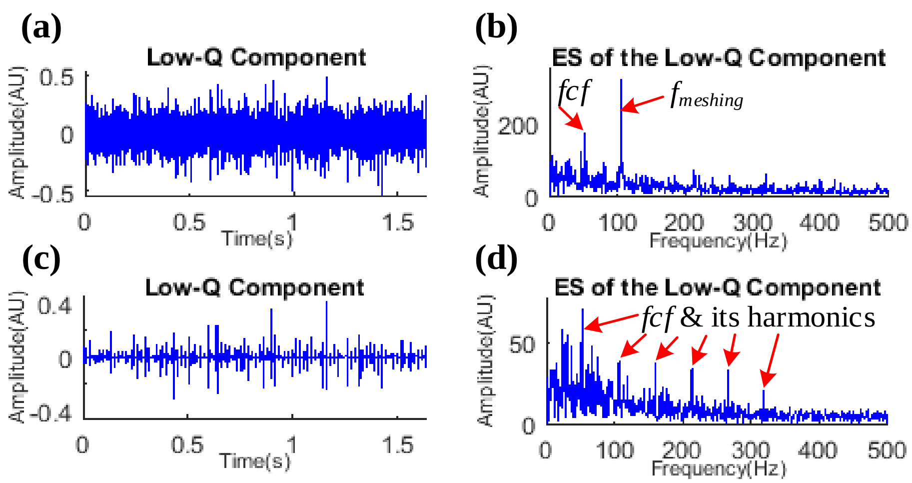

Fig. 3.7 Result of the OBSD with small regularization parameter ($\lambda_l=\lambda_h=0.01$) (a) and (b), and result with large regularization parameter ($\lambda_l=\lambda_h=0.3$) (c) and (d)

The objective function curves generated with different μ (μ=0.02 μ=0.05 μ=0.1 μ=0.5, μ=1, μ=2, μ=5 and μ=10) in the implementation of the OBSD are shown in Fig. 3.8. It can be seen that the convergence of the curve becomes faster with the increase of μ from 0.02 to 0.5. However, the convergence gets slower with further increase of μ from 0.5 to 10. Considering only the 8 given curves, the convergence rate reaches the highest value when μ=0.5. As mentioned previously, the convergence rate is affected by the μ but not in a monotonic increasing relationship. The simulations in Fig. 3.8 confirm the analysis.

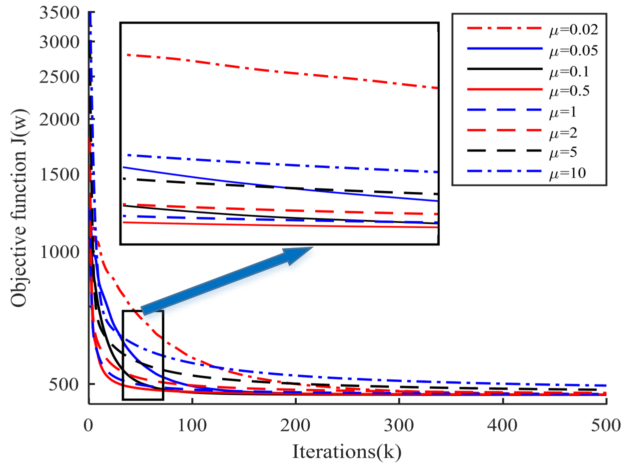

Fig. 3.8 Objective function curves with different penalty parameters (μ=0.02, μ=0.05, μ=0.1, μ=0.5, μ=1, μ=2, μ=5 and μ=10)

Finally, by reducing the number of iterations K to 50, the decomposed low-Q component via the OBSD and the corresponding envelope spectra are shown in Fig. 3.9(a) and Fig. 3.9(b), respectively. As seen in Fig. 3.9(b), the gear meshing frequency still shows the highest peak in the envelope spectrum of the low-Q component. From Fig. 3.8, it can be concluded that the objective function curve with μ=2 is not convergent enough after 50 iterations. It was discussed in the previous section that the choice of K has to ensure the objective function converges to a reasonable value. Otherwise, the result may not be reliable. The result in Fig. 3.9 verifies the analysis about the effects of the selection of K.

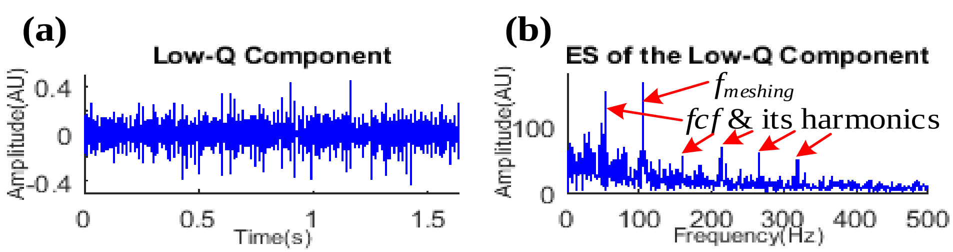

Fig. 3.9 Result of the OBSD with small number of iterations (K=50)

It is worth mentioning that the selection of μ is not needed to generate the fastest convergence rate. It can be seen that the results shown in Fig. 3.4(d) are good enough to detect the bearing fault, even though the μ is set as μ=2 instead of μ=0.5, which corresponds to the fastest convergence rate. The reason is that the number of iterations K is sufficient. As shown in Fig. 3.8, the values of the objective function with μ=2 and μ=0.5 are minimized to almost the same level after 200 iterations.

3.7 Conclusions

The OBSD can be extremely effective for bearing fault signature extraction in the presence of interference. The effects of the selection of the OBSD parameters on the performance of the method were analyzed in this chapter. Also, the results of the implementation of the OBSD on experimental signals validated the effectiveness of the analysis. Parameter selection for the application of the OBSD to bearing fault signature extraction can be carried out based on the discussed effects.

The development of automatic parameter selection algorithms for use with the OBSD is the topic of the next chapter.

Chapter 4 Auto-OBSD with an application to bearing fault signature extraction

This chapter addresses objective 2. An automatic parameter selection algorithm was proposed to improve the performance of the OBSD for the bearing fault signature extraction under constant speed condition.

The content of this chapter has been published in *Mechanical Systems and Signal Processing*.

Huan Huang, Natalie Baddour, and Ming Liang. "Auto-OBSD: Automatic parameter selection for reliable Oscillatory Behavior-based Signal Decomposition with an application to bearing fault signature extraction." *Mechanical Systems and Signal Processing* 86 (2017): 237-259.

4.1 Abstract

Bearing signals are often contaminated by in-band interferences and random noise. Oscillatory Behavior-based Signal Decomposition (OBSD) is a new technique which decomposes a signal according to its oscillatory behavior, rather than frequency or scale. Due to the low oscillatory transients of bearing fault-induced signals, the OBSD can be used to effectively extract bearing fault signatures from a blurred signal. However, the quality of the result highly relies on the selection of method-related parameters. Such parameters are often subjectively selected and a systematic approach has not been reported in the literature. As such, this chapter proposes a systematic approach to automatic selection of OBSD parameters for reliable extraction of bearing fault signatures. The OBSD utilizes the idea of Morphological Component Analysis (MCA) that optimally projects the original signal to low oscillatory wavelets and high oscillatory wavelets established via the Tunable Q-factor Wavelet Transform (TQWT). In this chapter, the effects of the selection of each parameter on the performance of the OBSD for bearing fault signature extraction are investigated. It is found that some method-related parameters can be fixed at certain values due to the nature of bearing fault-induced impulses. To adaptively tune the remaining parameters, index-guided parameter selection algorithms are proposed. A Convergence Index (CI) is proposed and a CI-guided self-tuning algorithm is developed to tune the convergence-related parameters, namely, penalty factor and number of iterations. Furthermore, a Smoothness Index (SI) is employed to measure the effectiveness of the extracted low oscillatory component (i.e. bearing fault signature). It is shown that a minimum SI implies an optimal result with respect to the adjustment of relevant parameters. Thus, two SI-guided automatic parameter selection algorithms are also developed to specify two other parameters, i.e., Q-factor of high-oscillatory wavelets and regularization parameter. Based on the index-guided parameter selection algorithms, an automatic parameter selection method is then proposed for the application of OBSD to bearing fault signature extraction. The proposed method is then validated with simulated signals and experimental data.

4.2 Introduction

Bearings are among the most frequently used components in rotating machines. Undetected bearing faults can increase in severity and lead to a breakdown of the machine and in some cases may lead to an accident. Therefore, bearing fault detection is very important and has been intensively investigated over the past few decades.

The high-frequency resonance technique is one of the most effective methods of bearing fault detection ([7],[9]) since the vibration signal excited by a local defect in the bearing has a relatively high damped resonance frequency. Therefore, plenty of research related to bearing fault detection has contributed to optimal band-pass filtering. For instance, the Spectral Kurtosis (SK) is found being able to indicate the bearing transient signal and hence to locate the frequency of the transient [23]. Then, the SK is further developed to design an optimal filter for bearing fault diagnosis [24]. A Smoothness Index (SI) is also proposed to guide the selection of the scale and the shape factor of the wavelet based de-noising algorithm [25], which can also be considered as the selection of center frequency and bandwidth of the optimal filter. Additionally, a resonance frequency estimation method has been developed by locating the frequency of maximum empirical probability density change to assist with optimal filtering [26]. The essence of band-pass based methods for bearing fault diagnosis is extracting the bearing fault signal in a specified band. Hence, these methods are sensitive to the frequencies of the signal components contained in the raw signal. Obviously, the effectiveness of these methods will be impaired if the frequency of the bearing fault signal and the frequency of any interference signals are similar or identical.

Recently, a resonance-based signal decomposition method, named Oscillatory Behavior-based Signal Decomposition (OBSD), was proposed by Selesnick [40]. Unlike band-pass based methods, the signal is decomposed according to the resonance or oscillatory behavior as defined by a Q-factor. Thus, it is independent of frequency features. This is particularly important in handling situations where the target frequency (e.g., bearing fault frequency) and interference frequencies both reside in the same frequency band. The Tunable Q-factor Wavelet Transform (TQWT) was developed to provide a set of over-complete bases which have the same oscillatory behavior as each other ([62],[88]). Since the bearing fault-induced signal is composed of impulsive responses which can be regarded as the low-oscillatory component with low Q-factor which can be easily separated from the high oscillatory ones, the application of the TQWT for bearing fault signal extraction has been explored recently ([64]–[66]). For example, a kurtosis-guided adaptive demodulation method based on the TQWT was proposed for bearing fault detection in [64]. A TQWT de-noising method with neighboring coefficients was also developed and applied to bearing fault diagnosis [65]. In addition, a fault feature ratio was proposed to choose proper Q-factors for the application of the TQWT to bearing fault feature extraction [66]. However, these aforementioned approaches still focus on locating the optimal sub-band of the TQWT for a bearing fault-induced signal. To take advantage of the resonance-based signal decomposition, the application of Morphological Component Analysis (MCA) based on the TQWT, which is the essence of the OBSD, is further investigated for bearing fault signature extraction. As the bearing fault signature is regarded as an impulsive low-oscillatory component whereas the interference is usually a smoother signal which can be considered as a high-oscillatory component, the bearing signal is decomposed into a low-oscillatory component (bearing transient signal) and a high-oscillatory component (interferences) by implementing the OBSD [67]. To reduce the influence of noise on the MCA, the Ensemble Empirical Mode Decomposition (EEMD) is jointly used with the OBSD [68]. Moreover, an iterative OBSD was developed to extract the bearing fault signature from a signal blurred by a series of interferences [69]. All these applications have shown that the OBSD can be used to effectively extract the bearing fault feature from the raw vibration signal contaminated by interferences and noise. However, there are a total of 10 parameters involved in the OBSD. Ill-selected parameters inevitably undermine the usefulness of the OBSD output and mislead fault detection decisions. Systematic OBSD parameter selection has not been adequately investigated in the literature. A subjective or trial-and-error approach is obviously not acceptable for quick and reliable applications. This study is therefore directed towards the development of a systematic, fast and automated approach to the proper selection of OBSD parameters for bearing fault signature extraction. The method is hereafter called Auto-OBSD.

The rest of this chapter is structured as follows. The motivation of this chapter is explained in section 4.3, where the typically-observed contaminated bearing signal and a short summary of the OBSD are introduced. The importance of parameter selection for application of the OBSD is then illustrated via two examples. In section 4.4, the OBSD is interpreted in detail, including the relationship between the Q-factor and the oscillatory behavior, the rationale of the TQWT, and the rationale of the MCA. Subsequently, the effects of parameter selection on the performance of the OBSD for bearing fault signature extraction are analyzed in section 4.5. In section 4.6, based on the prior analysis, three index-guided parameter selection algorithms are proposed to automatically select appropriate parameters. A Convergence Index (CI)-guided algorithm is proposed to ensure an acceptable level of convergence. Two Smoothness Index (SI)-guided algorithms are also proposed to warrant the effectiveness of the result. Based on the index-guided algorithms, an automatic parameter selection method is proposed in section 4.7 for the application of OBSD to bearing fault signature extraction. Additionally, the proposed automatic OBSD is validated by simulations. Experimental validation follows in section 4.8. Finally, conclusions are drawn in section 4.9.

4.3 Motivation

Bearing fault signature extraction is an important pre-processing step for bearing fault detection and fault diagnosis. The newly developed OBSD is a useful technique to extract the bearing fault feature from a signal obscured by interferences and noise. However, the OBSD is a parameter-dependent method. The parameters must be properly selected to ensure a reliable result.

4.3.1 Typical observed signals of defective bearings

The vibration signal acquired from a defective bearing is usually a mixture of the bearing fault-induced signal, interferences transmitted from other components and background noise, and can be written as

$$y(t) = x(t) + h(t) + n(t) \tag{4.1}$$

where $y(t)$ is the observed raw vibration, $x(t)$ is the pure bearing fault-induced signal, $h(t)$ represents interference components, and $n(t)$ is noise. The signal induced by a local bearing defect can be regarded as the impulse response of a one-degree-of-freedom mass-spring-damper system [9], and for a single event can be expressed as

$$s(t) = Ae^{-\beta t} \sin\left(\omega_r t\right) u(t) \tag{4.2}$$

where A is the amplitude, β is the damping coefficient, ω_r is the excited resonance frequency, and $u(t)$ is the unit step function. The impulse excitation is periodically applied at the Fault Characteristic Frequency (FCF). The FCF is proportional to the shaft rotational speed and can be determined by the structural parameters of the bearing [9]. Hence, the bearing fault-induced signal is modeled as [36]

$$x(t)=\sum_{m=1}^{M}A_m e^{-\beta\left(t-mT_p-\sum_{i=1}^{m}\tau_i\right)}\sin\left[\omega_r\left(t-mT_p-\sum_{i=1}^{m}\tau_i\right)\right]u\left(t-mT_p-\sum_{i=1}^{m}\tau_i\right) \tag{4.3}$$

where M is the total number of impulses which is determined by the signal length T and the value of the FCF (f_c), A_m is the amplitude of the m^{th} fault impulse response, T_P is the reciprocal of f_c, τ_i is the random slippage during each T_P, usually considered as $0.01T_P$~$0.02T_P$.

An example of the bearing fault signal is plotted in Fig. 4.1(a). From the magnified impulse response, it can be seen that the transient of the bearing fault-induced signal is impulsive. The interferences are signals that are transmitted from other sources such as gears. Interference components can be modelled as

$$h(t)=A_h g(t)=A_h\sum_{i=1}^{I}g_i(t)=A_h\sum_{i=1}^{I}\sum_{j=1}^{G_i}B_{ij}\sin\left(2\pi jf_{hi}\right) \tag{4.4}$$

where A_h is the amplitude of the interference signal, I is the total number of the interference components, G_i is the total number of harmonics of the i^{th} interference, B_{ij} is the amplitude ratio coefficient of the j^{th} harmonic of the i^{th} interference, f_{hi} is the fundamental frequency of the i^{th} interference. An example of the interference is shown in Fig. 4.1(b). It can be seen that the interference is smoother compared to the bearing fault-induced signal and oscillates continuously. Noise is usually modelled as random noise. The intensity of the noise can be measured by the signal-to-noise ratio (SNR). Similarly, the intensity of the interference can also be given by the signal-to-interference ratio (SIR). To unify the expression of the test signal, the contaminated bearing signals are defined by f_c, ω_r, β, τ_i, f_h, G, B_j, SIR and SNR throughout this chapter. Without loss of generality, the amplitude A_m is fixed to 1. If decibel (dB) is used as the unit of SIR [88], the amplitude of the interference signal A_h is then calculated as

$$A_h=10^{-\frac{SIR}{20}}\left(\frac{\int_0^T x^2(t)dt}{\int_0^T g^2(t)dt}\right)^{\frac{1}{2}} \tag{4.5}$$

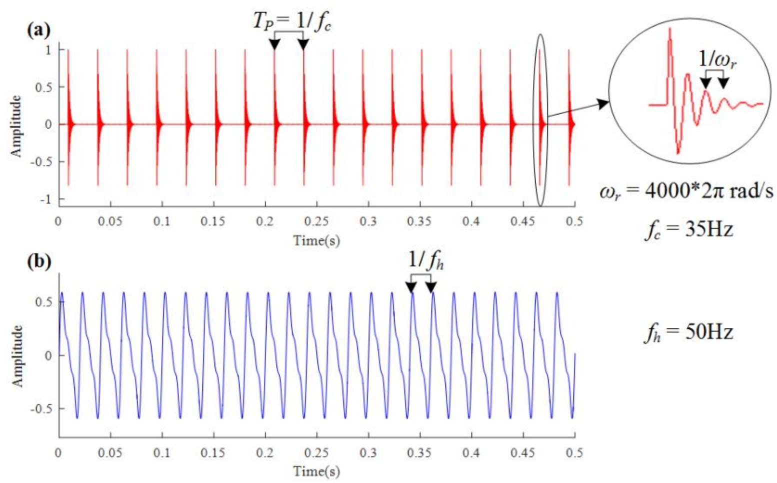

Fig. 4.1 (a) An example of the bearing fault-induced signal, and (b) an example of the interference

The bearing fault may not be correctly detected without signal extraction if the observed signal is contaminated by interferences. Bearing faults can be detected via the Hilbert envelope spectrum of the bearing signal [9]. If peaks appear at the FCF of a certain fault type in the envelope spectrum, the bearing is considered as faulty with the corresponding fault type [9]. The envelope spectrum is the frequency spectrum of the envelope of the signal. For a signal $x(t)$, the Hilbert envelope spectrum is defined as [42]

$$f(\omega) = F\left\{H\left[x(t)\right]\right\} = \int_{-\infty}^{+\infty}\left(\frac{1}{\pi}\int_{-\infty}^{+\infty} x(t)\frac{1}{t-\tau}d\tau\right)\cdot e^{-i\omega t}dt \qquad (4.6)$$

where $F\{\}$ is the operator of the Fourier Transform, and $H[]$ is the operator of the Hilbert Transform. An example of the envelope spectrum of a contaminated signal is shown in Fig. 4.2. The original signal is a mixture of the bearing fault-induced signal (f_c=35Hz) and two interference components (f_{h1}=36Hz and f_{h2}=500Hz), without noise. The relevant parameters used for the simulation of the mixed signal are ω_r=2000*2π rad/s, β=800, τ_i=0.01T_p, G_1=G_2=3, B_{1j}=B_{2j}=[2, 1, 0.5], SIR=-15dB. As shown in Fig. 4.2, the envelope spectrum is dominated by the frequencies of the interferences, their harmonics, and sidebands. The FCF and its multiples are totally submerged. It is obvious that the signal extraction is a necessary pre-processing step for bearing fault detection.

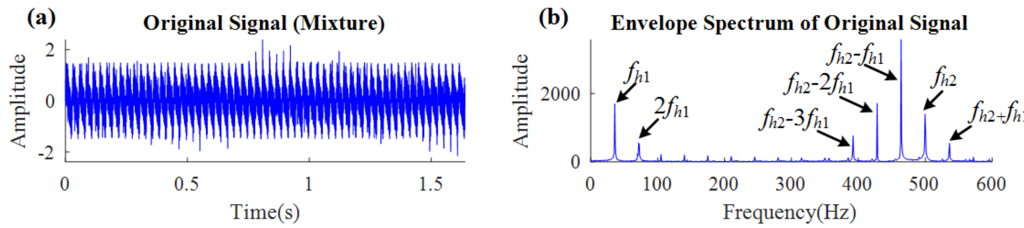

Fig. 4.2 A contaminated signal and its envelope spectrum

4.3.2 Short summary of the OBSD

Table 4.1 Steps of the Oscillatory Behavior-based Signal Decomposition (OBSD)

Step	Method	Input	Parameters	Output	Algorithms
1	TQWT (filter banks)	$\mathbf{y}$	Q_l, r_l, P_l Q_h, r_h, P_h	$\mathbf{S}_l, \mathbf{S}_h$	-
2	MCA	$\mathbf{y}$ $\mathbf{S}_l, \mathbf{S}_h$	λ_l, λ_h μ, K	$\mathbf{w}_l^{opt}$, $\mathbf{w}_h^{opt}$	Optimization: $\arg\min_{\mathbf{w}_l,\mathbf{w}_h} \lVert \mathbf{y}-\mathbf{S}_l\mathbf{w}_l-\mathbf{S}_h\mathbf{w}_h \rVert_2^2 + \lambda_l \lVert \mathbf{w}_l \rVert_1 + \lambda_h \lVert \mathbf{w}_h \rVert_1$ Solutions (via SALSA): $\mathbf{w}_i^k = f(\mathbf{w}_i^{k-1}, \mu)$, $i=l, h$; $k=1,2,\ldots,K$ $\mathbf{w}_i^{opt} = \mathbf{w}_i^K$, $i=l, h$
3	iTQWT	$\mathbf{w}_l^{opt}$, $\mathbf{w}_h^{opt}$ $\mathbf{S}_l, \mathbf{S}_h$	-	$\hat{\mathbf{y}}_l$, $\hat{\mathbf{y}}_h$	$\hat{\mathbf{y}}_l = \mathbf{S}_l\mathbf{w}_l^{opt}$, $\hat{\mathbf{y}}_h = \mathbf{S}_h\mathbf{w}_h^{opt}$

The OBSD, developed by Selesnick, aims to separate a signal into two components which have different oscillatory behaviors (or resonance characters) in sparse representations [40]. The steps of this novel signal decomposition method are given in Table 4.1. The representations of the involved parameters are given in Table 4.2. A complex non-stationary signal $\mathbf{y}$ can be considered as a combination of a component consisting of low oscillatory behavior $\mathbf{y}_l$ and a component consisting of high oscillatory behavior $\mathbf{y}_h$. By applying the OBSD, the components are estimated as $\hat{\mathbf{y}}_l$ and $\hat{\mathbf{y}}_h$, respectively. As shown in Table 4.1, there are three steps to accomplish the OBSD. First, an Over-Complete Dictionary (OCD) is built up via the filter banks of the TQWT. The OCD is composed of the oscillatory wavelets $\mathbf{S}_l$ and high oscillatory wavelets $\mathbf{S}_h$. Then, the MCA is employed to sparsely represent the low and high oscillatory components with wavelet coefficients $\mathbf{w}_l^{opt}$ and $\mathbf{w}_h^{opt}$, respectively. The MCA is essentially an optimization algorithm devoted to determining the optimal wavelet coefficients. To solve the optimization problem, the Split Augmented Lagrangian Shrinkage Algorithm (SALSA) is implemented [86], which is a special case of the Alternating Direction Method of Multipliers (ADMM) [89]. Once the optimal wavelet coefficients are obtained, the components of high and low oscillation are then determined via inverse TQWT. All the related algorithms are explained in detail in section 4.4.

Table 4.2 Parameters in Table 4.1

Parameter	Representation	Parameter	Representation
$\hat{\mathbf{y}}_l$	Estimated low-oscillatory component	$\hat{\mathbf{y}}_h$	Estimated high-oscillatory component
$\mathbf{S}_l$	Low-oscillatory wavelets	$\mathbf{S}_h$	High-oscillatory wavelets
Q_l	Q-factor of $\mathbf{S}_l$	Q_h	Q-factor of $\mathbf{S}_h$
r_l	Redundancy of $\mathbf{S}_l$	r_h	Redundancy of $\mathbf{S}_h$
P_l	Number of levels of $\mathbf{S}_l$	P_h	Number of levels of $\mathbf{S}_h$
λ_l	Regularization parameter of $\hat{\mathbf{y}}_l$	λ_h	Regularization parameter of $\hat{\mathbf{y}}_h$
μ	Penalty factor	K	Number of iterations
$\mathbf{w}_l^k$	Wavelet coefficients of kth iteration for $\hat{\mathbf{y}}_l$	$\mathbf{w}_h^k$	Wavelet coefficients of kth iteration for $\hat{\mathbf{y}}_h$
$\mathbf{w}_l^{opt}$	Optimal wavelet coefficients for $\hat{\mathbf{y}}_l$	$\mathbf{w}_h^{opt}$	Optimal wavelet coefficients for $\hat{\mathbf{y}}_h$

4.3.3 Implementation examples of the OBSD for bearing fault signature extraction

The contaminated bearing signal $y(t)$ can be considered as the input $\mathbf{y}$ of the OBSD. The bearing fault-generated signal $x(t)$ can be considered as the low-oscillatory component $\mathbf{y}_l$ of the total signal, whereas the interference $h(t)$ is considered to be the high-oscillatory component $\mathbf{y}_h$, both extracted via the OBSD. According to Fig. 4.1, the bearing fault-induced signal has fewer oscillations compared with the interference. However, as shown in Table 4.1, ten parameters, i.e., Q_l, r_l, P_l, Q_h, r_h, P_h, λ_h, λ_l, K and μ, have to be selected when implementing the OBSD.

If the parameters are properly selected, the bearing fault signature can be well separated from the mixture. An example of the application of the OBSD to the bearing signal extraction is shown in Fig. 4.3. The original signal is the same signal shown in Fig. 4.2(a). The OBSD-related parameters are set as Q_l=1, r_l=6, P_l=45, Q_h=6, r_h=6, P_h=144, λ_h=0.3, λ_l=0.3, K =200 and μ=2. The decomposed low-oscillatory component, high-oscillatory component and residual are shown in Fig. 4.3(c), Fig. 4.3(e), and Fig. 4.3(g), respectively. In addition, the envelope spectrum of the low-oscillatory component and the envelope spectrum of the high-oscillatory component are shown in Fig. 4.3(d) and Fig. 4.3(f). It can be seen that the FCF and its harmonics cannot be identified from the Hilbert envelope spectrum of the original signal. Thus, the bearing fault cannot be detected. However, by applying the OBSD, the mixture of the impulsive bearing signal and the oscillatory interference is decomposed into a low-oscillatory component and a high-oscillatory component. The location of the impulses in the low-oscillatory component is exactly the same as in the pure bearing fault signal. In addition, the interference is completely absent from the low-oscillatory component. The same result is revealed in the envelope spectrum of the low-oscillatory component, namely that the FCF and its harmonics take over the dominant position from the harmonic frequencies of the interference.

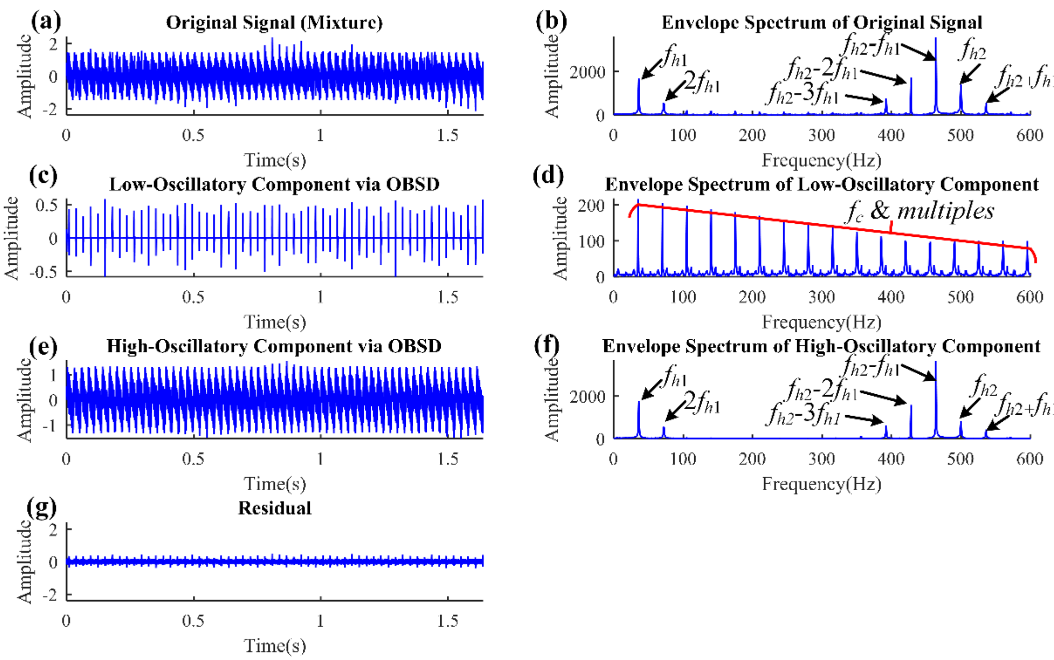

Fig. 4.3 Implementation of the OBSD on bearing fault signature extraction with proper parameters

However, the bearing fault features may not be fully extracted if the parameters are improperly selected. Another example of the application of the OBSD to the bearing signal decomposition with the same test signal is shown in Fig. 4.4. The parameters are kept the same, with the exception of three modifications: Q_h=2, λ=0.1, and P_h=67. The result shows that the OBSD has failed almost completely: the frequency spectrum of the low-oscillatory component is still dominated by the two interferences and one can hardly detect any trace of the FCF and its harmonics in the Hilbert envelope spectrum of the low-oscillatory component (Fig. 4.4(d)). It clearly demonstrates the importance of parameter selection for reliable bearing fault signature extraction via the OBSD.

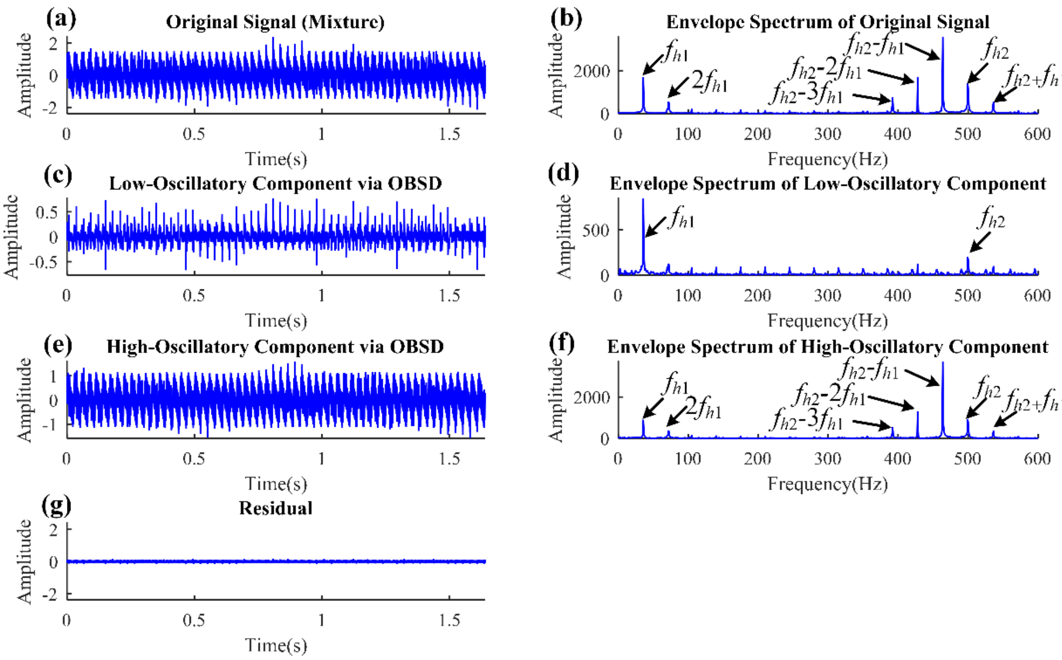

Fig. 4.4 Implementation of the OBSD on bearing fault signature extraction with improper parameters

4.4 Oscillatory Behavior-based Signal Decomposition (OBSD)

The OBSD method is developed based on the MCA idea that separates a signal into components of different morphological features, such as different oscillatory behaviors. The TQWT is employed to assist the MCA for the OBSD since it is available to generate wavelet bases according to specified oscillatory levels.

4.4.1 Q-factor and oscillatory behavior

The Q-factor is defined as the ratio of the center frequency to the bandwidth of a band-pass filter [62]. It is believed that the Q-factors of a waveform in different scales are the same [90]. Furthermore, the frequency response of a waveform with lower oscillatory behavior has a lower Q-factor whereas a waveform with high oscillatory behavior has a higher Q-factor. An example of the Daubechies wavelets in different scales and the corresponding frequency spectra are shown in Fig. 4.5, which is cited from [40] with modifications. Waveform 1 and waveform 3 are variants of the same wavelet, which means that they have the same oscillatory behavior. From the frequency spectrum, it can be seen that the Q-factors of the two waveforms are both equal to 1. Similarly, waveform 2 and waveform 4 are also variants of the same wavelet.

The Q-factors of these two waveforms are the same, both equal to 5. Therefore, the oscillatory level of a signal can be described by its Q-factor. It can be also seen that waveform 1 and waveform 2 are at the same frequency since the corresponding center frequencies are the same. Similarly, waveform 3 and waveform 4 are at the same frequency as well. It is obvious that different signals with similar frequencies cannot be separated via frequency-based filtering methods. However, the Q-factor can be used to identify the signals with different oscillatory behaviors even when they are at the same frequencies.

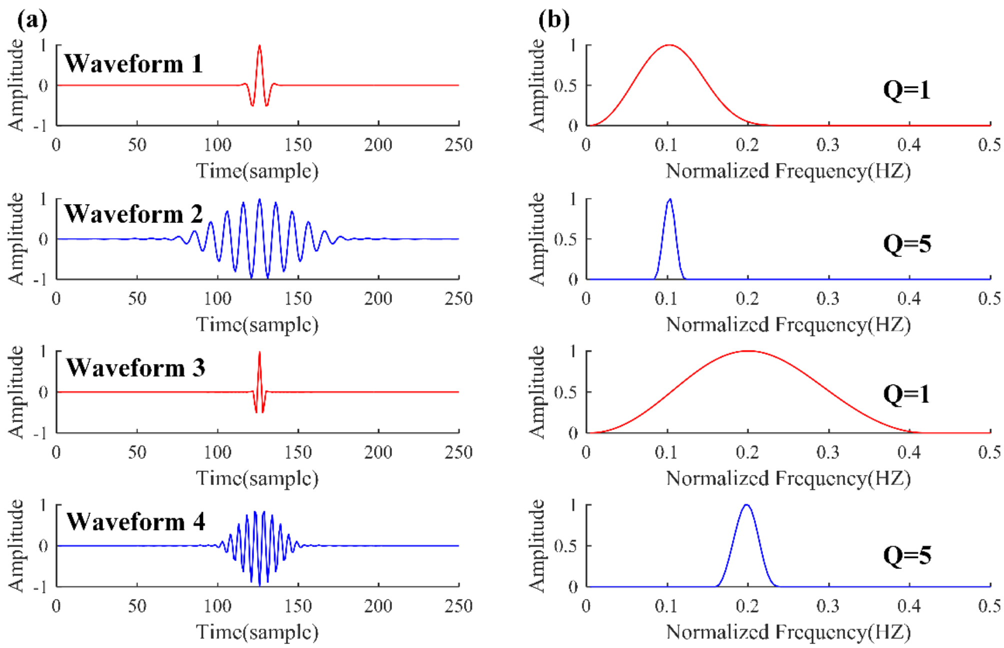

Fig. 4.5 Waveforms and their frequency spectra (cited from [40] with modifications)

4.4.2 TQWT and sparse signal representation

The TQWT is a constant Q-factor Discrete Wavelet Transform (DWT) with over-complete bases developed by Selesnick [62]. The Q-factor can be continuously adjusted to match the oscillatory behavior of the analyzed signal.

First, the TWQT is a constant Q-factor DWT which means the wavelets (i.e. the basis functions) have the same Q-factor. The structure of the TQWT is based on the discrete dyadic DWT which is a two-channel filter bank. As shown in Fig. 4.6, the filter bank is composed of a number of stages. For each stage, it consists of two channels which are a low-pass filter with frequency response $H_0(\omega)$ and a high-pass filter with frequency response $H_1(\omega)$. The outputs of the filters are also re-scaled by a Low-Pass Scaling with parameter α (LPS α) and a High-Pass Scaling with parameter β (HPS β), respectively. The samples of the high-pass signals then constitute the wavelet coefficients.

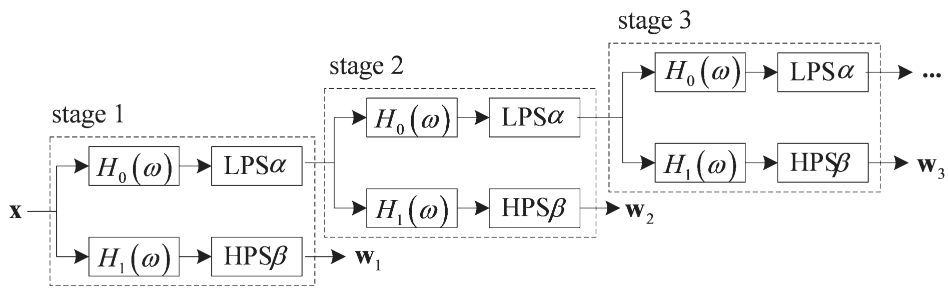

Fig. 4.6 Two-channel filter banks of the TQWT

In order to achieve the sparse signal representation, the over-complete bases are constructed with redundancy r. Compared with orthogonal bases, over-complete bases are more flexible in their ability to represent the signal sparsely [63]. By implementing a basis pursuit or other similar optimization algorithm, optimal sparse wavelet coefficients can be obtained. In addition, the TQWT is also designed for perfect reconstruction.

Once the Q-factor is selected, the waveform of the bases for the TQWT is determined. By choosing the redundancy r and the total level of the stages P, the frequency responses of the filter banks are determined. Note that the total level of the wavelet coefficients is P+1 (P outputs of P high-pass filters plus the output of the low-pass filter of the P^{th} stage). The basis functions (wavelets) and the frequency response of the TQWT are shown in Fig. 4.7 when the parameters are set as Q=3, r=3, and P=15. Details about the implementation of the TQWT and the inverse TQWT can be found in Appendix A.

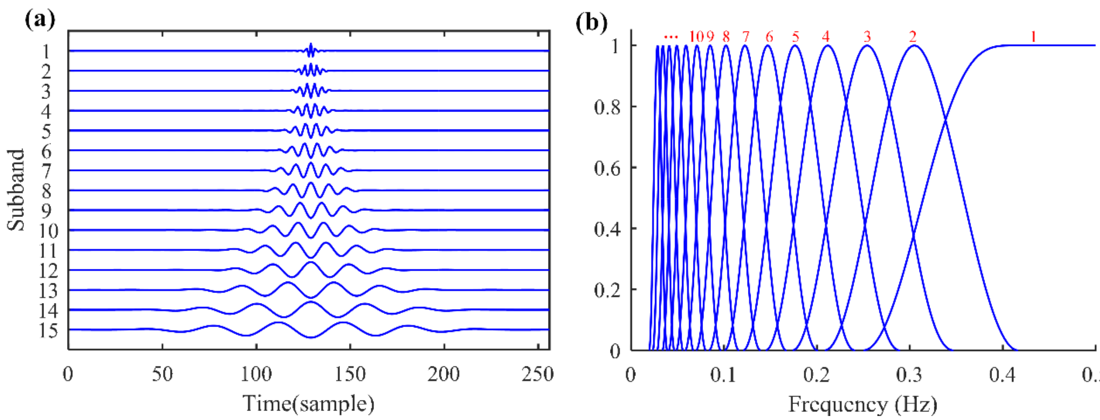

Fig. 4.7 Wavelets of each sub-band and the corresponding frequency response of TQWT when Q=3, r=3, and P=15

4.4.3 Morphological Component Analysis (MCA)

The MCA method was originally proposed to separate features which present different morphological aspects in an image [91]. As the oscillatory behavior of a signal could be regarded as the morphological features, the MCA method was applied to decompose the vibration signal. The goal of the MCA is to determine the optimal sparse representations of the low oscillatory component $\mathbf{w}_l^{opt}$ and those of the high oscillatory component $\mathbf{w}_h^{opt}$ separately.

The optimal coefficients can be considered as an optimization problem with the objective function given as [40]

$$J\left(\mathbf{w}_l, \mathbf{w}_h\right) = \left\|\mathbf{y} - \mathbf{S}_l \mathbf{w}_l - \mathbf{S}_h \mathbf{w}_h\right\|_2^2 + \lambda_l \left\|\mathbf{w}_l\right\|_1 + \lambda_h \left\|\mathbf{w}_h\right\|_1 \tag{4.7}$$

where $J(\mathbf{w})$ is the objective function, $\mathbf{w}_l$ are the wavelet coefficients of the low-oscillatory component, and $\mathbf{w}_h$ are the wavelet coefficients of the high-oscillatory component. The interpretation of the other variables can be found in Table 4.2. The objective function is the sum of the l_2-norm squared of the residual and l_1-norm of $\mathbf{w}_l$ and $\mathbf{w}_h$. By minimizing the objective function, the optimal wavelet coefficients can be obtained from

$$\begin{bmatrix} \mathbf{w}_h^{opt} \\ \mathbf{w}_l^{opt} \end{bmatrix} = arg \min_{\mathbf{w}_h, \mathbf{w}_l} \left\|\mathbf{y} - \mathbf{S}_h \mathbf{w}_h - \mathbf{S}_l \mathbf{w}_l\right\|_2^2 + \lambda_h \left\|\mathbf{w}_h\right\|_1 + \lambda_l \left\|\mathbf{w}_l\right\|_1 \tag{4.8}$$

The optimal solution has to satisfy two conditions which are a sparse representation and a small residual. The regularization parameters λ_h and λ_l can be used to balance the l_2-norm term and the l_1-norm term. If λ_h and λ_l are large, sparsity of $\mathbf{w}_l$ and $\mathbf{w}_h$ carry more influence than the energy of the residual. The simplest case is that the solution is only constrained by the first term, which is represented as λ_h=λ_l=0. Moreover, λ_h and λ_l are the measurement of the intensity of the sparse representation of $\mathbf{w}_l$ and $\mathbf{w}_h$, respectively. This unconstrained optimization problem can be solved by the Split Augmented Lagrangian Shrinkage Algorithm (SALSA) as [40]

$$\mathbf{w}_i^k = f(\mathbf{w}_i^{k-1}, \mu),\ i = l, h,\ k = 1, 2, 3..., K \tag{4.9}$$

The details of the SALSA can be found in Appendix B. When the maximum number of iterations K is reached, the wavelet coefficients for the low and high oscillatory components are given as

$$\mathbf{w}_i^{opt} = \mathbf{w}_i^K, i = h, l \tag{4.10}$$

Once the optimal coefficients are acquired, the inverse TQWT (iTQWT) is applied to get the corresponding estimated components $\hat{\mathbf{x}}_h$ and $\hat{\mathbf{x}}_l$ from

$$\hat{\mathbf{y}}_i = \mathbf{S}_i \mathbf{w}_i^{opt}, i = h, l \tag{4.11}$$

4.5 Effects of parameter selection on OBSD for bearing fault feature extraction

There are ten parameters involved in the OBSD which are listed in Table 4.2. They can be classified as parameters related to the TQWT and parameters related to MCA as detailed below.

4.5.1 Parameters related to the TQWT

The role of the TQWT in the OBSD is to provide two sets of over-complete bases to estimate the bearing fault-induced signal and the interference signal specifically for bearing fault signature extraction. The establishment of the bases relies on the selection of Q, r, and P.

Thus, there are 6 parameters needed to be selected for the TQWT which are Q_l, r_l, and P_l for the low-oscillatory bases, and Q_h, r_h, and P_h for the high-oscillatory bases.

Effects and selection guides of the parameters related to the TQWT for bearing signature extraction is given in Table 4.3. Effects of the parameters including the overlapping extent of the frequency responses of the TQWT and the computational cost are provided.

Table 4.3 Effects and selection guides of the parameters related to the TQWT for bearing fault signature extraction

Parameter	overlapping extent of frequency responses	Computational cost	Selection guide
P	-	↑ if P ↑ [62]	$P_i=P_{imax}$, $i=l$, h
Q	-	↓ if Q ↑ when $P=\infty$ [62]	$Q_l=Q_{\min}=1$
			Q_h=oscillatory level of the interference
r	↑ if r ↑	↑ if r ↑ when $P=\infty$ [62]	r_i=trade-off(overlapping extent of frequency responses vs computational cost), $i=l$, h
Note: “-” represents no direct effect			

The number of stages of the filter banks, P, reveals the coverage of the frequency responses in the frequency domain. The computational cost increases with an increase in P. However, the P should be selected to be as large as possible, as given by Eq.(4.12) below [62]. Otherwise, the frequency responses may not cover the low frequencies. For example, if P is chosen as 10 for the case presented in Fig. 4.7, the approximate normalized frequency range (0, 0.08) is not handled by any of the high-pass filters. This will lead the TQWT back to a frequency based method instead of a fully oscillatory behavior based method. Therefore, if the frequency of the interference is not known prior, then P should always be chosen as the maximum possible value. The maximum of P is given by

$$P_{\max}=\left\lfloor \frac{\log\left(\frac{N}{4(Q+1)}\right)}{\log\left(\frac{(Q+1)r}{(Q+1)r-2}\right)} \right\rfloor \tag{4.12}$$

where $\lfloor A \rfloor$ rounds a number A to the next smaller integer, and N is the length of the signal. It can be seen that P_{max} is thus determined by Q, r, and N, which implies that the selection of P can be automatically completed once the Q-factor and redundancy are confirmed.

The Q-factor is used to describe the oscillatory level of the wavelet bases. To decompose the signal as expected, the Q-factors should be chosen to match the oscillatory levels of the target signal. Ideally, the Q-factor of the low-oscillatory bases Q_l should be selected as low as possible due to the impulsiveness of the bearing fault-induced signal. Because the Q-factor cannot be less than 1 it is thus reasonable to fix Q_l to 1 for bearing signal decomposition. Meanwhile, Q_h should be chosen to adapt to the oscillatory features of the interference. Additionally, as given in [62], the computational cost of the TQWT is $O(3rN\log_2(4N/(Q+1)))$ as P approaches ∞. This implies that the computational cost reduces with an increase in Q when $P=P_{\max}$. Also, in order to minimize the coherence of the low oscillatory wavelets and the high oscillatory wavelets, Q_h should be selected to be sufficiently greater than Q_l [40]. However, if Q_h is too high, the oscillatory behavior may not well match the expected high oscillatory component. This highlights the need for a method to determine appropriate values of Q_h that preserve the essential features of the signal, as shall be presented in section 4.6.3.

The redundancy r is related to the extent to which the frequency responses of the transform overlap. If r is increased, the density of frequency responses will be raised. However, the computational cost of the TQWT is linear in r [62]. Hence, an increase in r boosts the extent of overlapping of frequency responses, but lowers computational efficiency.

4.5.2 Parameters related to the MCA

The goal of the MCA method is to obtain the optimal wavelet coefficients by solving the optimization problem. There are four parameters related to the MCA, the two regularization parameters λ_l and λ_h involved in the objective function, the penalty factor μ, and the number of iterations K related to the SALSA.

The selection guide for the parameters related to the MCA for bearing signal extraction is given in Table 4.4. The parameters affect the speed of convergence of the solution, the value of the objective function $J(\mathbf{w}^k)$ after a specified number of iterations K, the minimum of the objective function $J_{min}(\mathbf{w})$, the sparsity of the wavelet coefficients and the computational cost. The selection of each parameter can be made based on the aforementioned properties.

Table 4.4 Effects and selection guide of the parameters related to the MCA for bearing signal extraction

Parameter	Convergence rate	$J(\mathbf{w}^K)$	$J_{min}(\mathbf{w})$	Sparsity	Computational cost	Selection guide
K	-	↓ if K ↑	-	-	↑ if K ↑	K=trade-off ($J_K(\mathbf{w})$ vs computational cost)
μ	↓ if μ ↑ when μ is not too small	-	-	-	-	μ for fast convergence rate
λ_l	-	-	↓ if λ_l ↑	↑ if λ_l ↑	-	λ_l for proper sparsity of $\mathbf{w}_l^{opt}$
λ_h	-	-	↓ if λ_h ↑	↑ if λ_h ↑	-	$\lambda_h = \lambda_l$
Note: “-” represents no direct effect						

The number of iterations, K, must ensure that the objective function $J(\mathbf{w}^K)$ reaches an acceptable level of convergence. Otherwise, the result may not be reliable. Theoretically, more iterations give a higher level of convergence. However, the computational cost is also raised with an increase in K. Therefore, the selection of K must be made to balance the convergence level of the result and the computational efficiency.

The penalty factor, μ, is related to the convergence speed of the solution. The parameter μ should be selected as one that generates the fastest convergence rate. However, the relation between μ and the convergence speed is not monotonic. In [86], it was found that the value of μ for fastest convergence differs for different cases. By applying the OBSD to bearing fault signature extraction with a series of μ, the $J(\mathbf{w})$-k curves are shown in Fig. 4.8(a). All curves are convergent to the minimum objective. A local version of the curves within iterations [40, 60] is shown in Fig. 4.8(b). It can be seen in Fig. 4.8(b) that the curve generated with μ=0.02 is the highest. When μ increases to 0.05, 0.1 and 0.5, the curve becomes lower and lower. Also, the curve associated with μ=0.5 is the lowest, and it reaches its steady state earlier than any other curves. However, when μ further increases to 1, 2, 5 and 10, the trend reverses, i.e., further increasing μ value leads to higher curve. In other words, the convergence rate increases when μ varies from 0.02 to 0.5, and then decreases with the growth of μ from 0.5 to 10. Therefore, if μ is too small or too large, the convergence of the algorithm may be slow (more iterations needed for convergence).

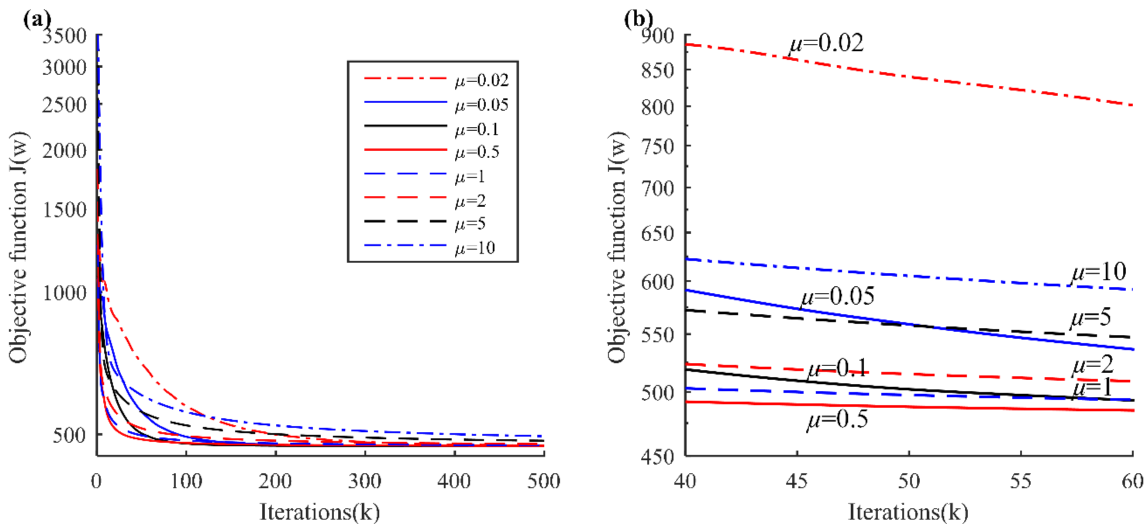

Fig. 4.8 (a) $J(\mathbf{w})$-k curves by applying the OBSD with a series of μ, and (b)local version of (a)

The regularization parameters λ_l and λ_h are used to balance the residual and the sparsity. As mentioned in section 4.4.3, with an increase of λ_l and λ_h, sparsity becomes more important than the small residual. In such a case, the obtained wavelet coefficients are sparser. However, the optimal solution may be unreliable if the wavelets coefficients are too sparse. The λ_l and λ_h also control the relative sparsity between the low-oscillatory wavelet coefficients and high-oscillatory wavelet coefficients. For the application to bearing fault detection, only the decomposed low-oscillatory component is further analyzed for the detection of bearing faults. Thus, it is not necessary to pay much attention to the relative sparsity. In conclusion, the λ_h and λ_l can be taken as λ_h=λ_l= λ.

4.6 Index guided parameter selection algorithms

The realization of automatic parameter selection algorithms can effectively improve the practicability of the OBSD. Based on the previous analysis, Q_l can be fixed to 1 and J is dependent on Q and r. If r_l and r_h can be also fixed, there are four parameters left to be tuned, Q_h, λ, K, and μ. These four parameters are mainly related to two important conditions that the result has to satisfy. First, the result is obtained under a desirable degree of convergence. Second, the decomposed low-oscillatory component should resemble the target bearing fault signature. To address the first problem, a Convergence Index (CI) is introduced and a CI-guided parameter selection algorithm is proposed to automatically tune the convergence-related parameters K and μ. In addition, inspired by the method reported in [25] a Smoothness Index (SI) is introduced to indicate the similarity of the extracted low-oscillatory component to the bearing fault feature. It is found the SI decreases with the reduction of the interference yet increases in the absence of fault-generated impulses. Based on this observation, two SI-guided parameter selection algorithms are developed to locate the proper λ and Q_h by minimizing the associated smoothness indices.

4.6.1 Convergence index guided selection of penalty factor and iteration number

The convergent properties of the OBSD are related to μ and K, as previously discussed. The selection of K must ensure that $J(\mathbf{w}^K)$ achieves convergence. To test the convergence of the result, a Convergent Index (CI) is proposed as

$$CI\left(J\left(\mathbf{w}^k\right)\right)=\frac{\left|J\left(\mathbf{w}^k\right)-J\left(\mathbf{w}^{k-1}\right)\right|}{J\left(\mathbf{w}^k\right)} \tag{4.13}$$

The K can be selected such that $J(\mathbf{w}^K)$ reaches an acceptable degree of convergence such that CI=ε. However, due to an improper rate of convergence, the K might be too large or too small even if $J(\mathbf{w}^K)$ is settled. If K is too large, the computational cost will be too high. Meanwhile, if K is too small, the result may not be reliable. Thus, μ must also be tuned to maintain a suitable convergence speed.

The proposed CI-guided self-tuning algorithm for K and μ is shown in the flow chart of Fig. 4.9. The conditions for the selection are given as

$$\mu, K = s.t. \begin{cases} CI\left(J\left(\mathbf{w}^{K}\right)\right) < \varepsilon \\ K \in \left[K_{\min}, K_{\max}\right] \\ \mu \geq \mu_{\min} \end{cases} \tag{4.14}$$

The convergence limit ε is set as 0.0001 in this chapter. To control the rate of convergence, K is restricted to $[K_{min}, K_{max}]$ and μ is limited as $\mu \geq \mu_{min}$. Once the convergence criterion is satisfied and K is located within the range $[K_{min}, K_{max}]$, the selection of μ is regarded as suitable. To ensure the convergence rate increases with the reduction of μ, μ_{min} is restricted to 0.5. Thus, if K exceeds the upper boundary K_{max}, the rate of convergence must be enhanced which means μ must be reduced. On the other hand, μ has to be increased if K is less than K_{min}. Considering the computational complexity and the reliability of the result, the range of K is fixed to be within [50,300]. The initial μ is set to 1 and the adjustment step $\Delta\mu$ is set to 0.5, respectively.

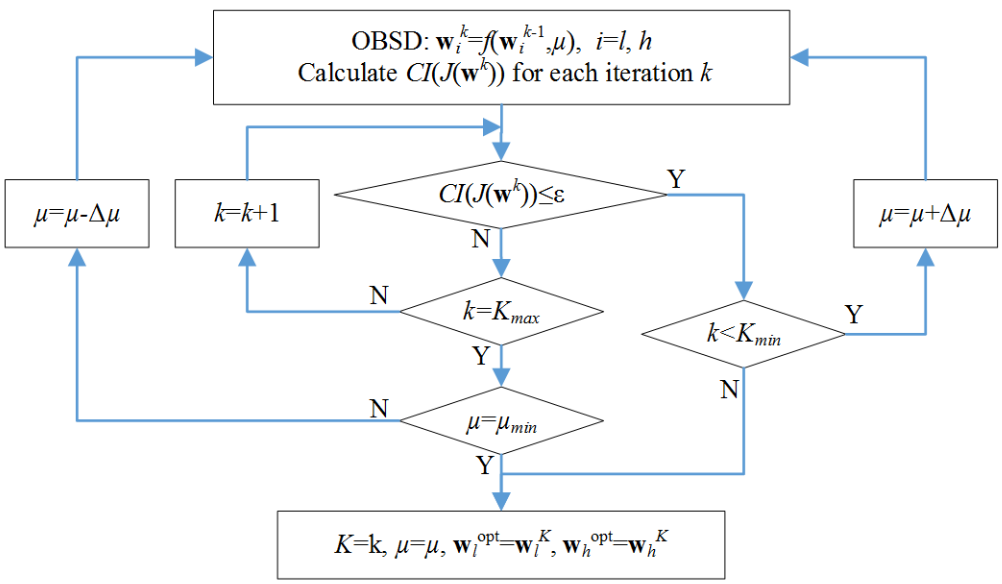

Fig. 4.9 Flowchart of CI-guided self-tuning algorithm for μ and K

4.6.2 Smoothness index guided regularization parameter selection

In addition to convergence related parameters, there are three other parameters, Q_h, λ_h, and λ_l, which critically affect the results. Since we take $\lambda_h = \lambda_l = \lambda$ in this chapter, there are only two parameters left. The decomposed low-oscillatory component via the OBSD is considered to be the bearing fault-induced signal. If the quality of the decomposed signal can be measured by an index, automatic parameter selection is feasible.

A Smoothness Index (SI) is introduced to indicate the smoothness of the result in this chapter. The SI of a discrete signal is defined as the ratio of the geometric mean of the absolute value of the signal and the arithmetic mean of the absolute value of the signal [25]. For an *N*-point signal *X*, the SI is expressed as

$$SI\left(X\right)=\frac{\left(\prod_{i=1}^{N}\left|X_i\right|\right)^{1/N}}{\left(\sum_{i=1}^{N}\left|X_i\right|\right)\Big/N} \tag{4.15}$$

The rationale of using SI is, as indicated in [25], that it “approaches unity for flat functions and zero for peaky functions”. Since it is known that the bearing fault-induced signal is highly impulsive, such a well-processed signal should have a very low degree of smoothness. Following the same logic, the SI of an appropriately decomposed low-oscillatory component should also be low and therefore the SI can be used to guide the selection of λ. As mentioned in previous sections, the selection of λ affects the sparsity of the result. Fig. 4.10 shows the decomposed low-oscillatory component via the OBSD as a function of increasing λ. The mixed signal is the same as the one used in section 4.3.3. It can be seen that with the growth of λ, the separated low-oscillatory component becomes sparser. At the beginning, the interference is increasingly removed from the low-oscillatory component when λ is increased. The SI of the low-oscillatory component declines correspondingly. The interference is fully removed when λ rises to a specified level, such as λ=0.3 in the given example. However, when λ continues to increase, the decomposed signal becomes too sparse to show the character of the impulses. Fortunately, the SI climbs rather than drops when the result becomes too sparse. The minimum SI thus points to the desired result. Therefore, the SI is suitable to indicate the appropriateness of λ.

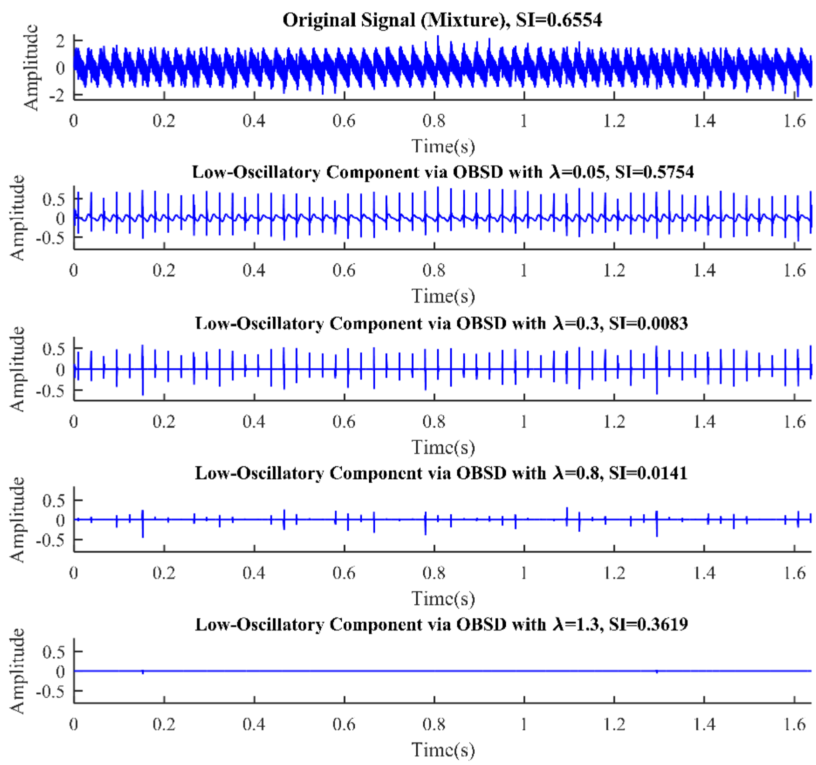

Fig. 4.10 Decomposed low-oscillatory component via the OBSD with increasing λ

Based on the observed relationship between the minimum SI and the proper λ, the λ can be tuned by locating the decomposed low-oscillatory component with the minimum SI, expressed as

$$\lambda = \arg\min_{\lambda} SI\left(\hat{\mathbf{y}}_l\right) \tag{4.16}$$

An optimal λ-searching algorithm is proposed in Fig. 4.11. The searching of proper λ departs from an initial λ_0, which is fixed at 0.5 in this chapter, and moves forward with a step $\Delta\lambda$ set as 0.1. The search direction, i.e. increasing or decreasing the λ, is determined by comparing the SI of the results obtained for $\lambda = \lambda_0 + \Delta\lambda = 0.6$ and $\lambda = \lambda_0 - \Delta\lambda = 0.4$ with the initial result. The process continues until the SI stops decreasing. The optimal λ is then considered to be located where the SI of the low-oscillatory component is the smallest throughout the searching process.

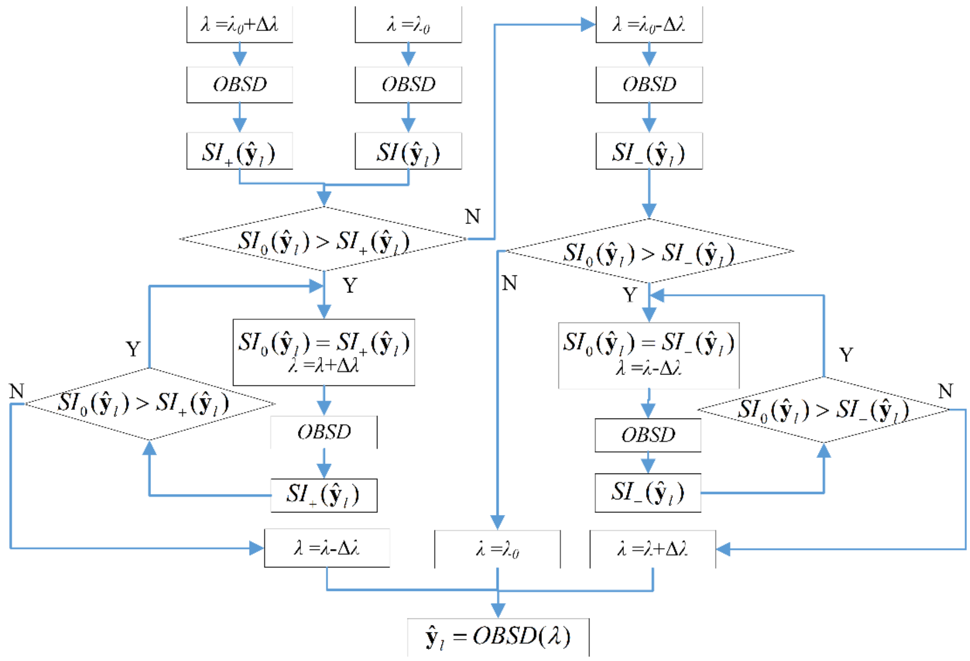

Fig. 4.11 Flowchart of SI-guided self-tuning algorithm for λ

It is worth mentioning that the SI is superior to kurtosis or entropy as a guide to the selection of λ. A graph composed of curves of SI vs λ, Kurtosis vs λ and Entropy vs λ is given in Fig. 4.12. To properly show the changes of all curves, the amplitude of each curve is normalized by setting the maximum value of each curve to be 1. Kurtosis is widely implemented to test the impulsiveness of a signal ([9], [23], [24]). For an N-point signal X, kurtosis is defined as

$$Kurtosis\left(X\right)=\frac{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^4/N}{\left(\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2/N\right)^2} \tag{4.17}$$

The maximum kurtosis can be used in various applications to indicate some kind of optimal result ([24],[64]). However, for the OBSD, maximum kurtosis does not point to an optimal result. As shown in Fig. 4.12, the kurtosis increases with the cancellation of the interference. However, it keeps rising even when the signal becomes too sparse to represent any bearing fault features. Furthermore, as observed by Bozchalooi and Liang in [25], the smoothness index, as a signal processing criterion, is much more robust than kurtosis in dealing with outliers in a signal. Shannon entropy can be also used to measure the randomness of a signal, defined as [92]

$$H(X) = -\sum_{i=0}^{n} p(x_i)\log_2 p(x_i) \tag{4.18}$$

where $p(x_i)$ is the probability of the i^{th} symbol of the signal X. As can be seen in Fig. 4.12, entropy decreases with an increase in λ, implying that it cannot used to indicate the selection of λ.

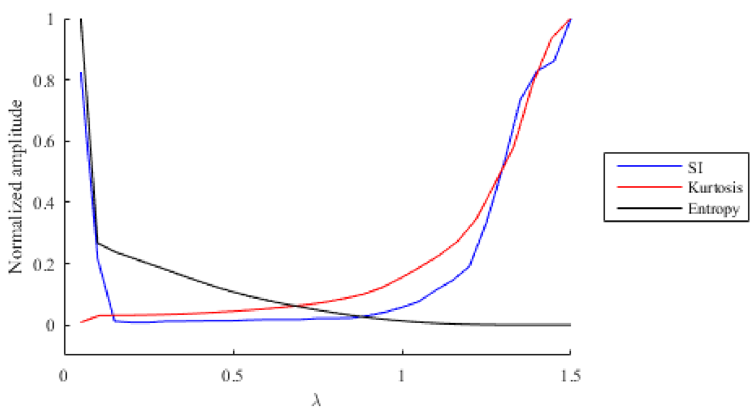

Fig. 4.12 Curves of SI vs λ, Kurtosis vs λ and Entropy vs λ

4.6.3 Smoothness index guided parameter selection for high Q-factor

The advantages of the SI as described above suggest that SI can also be applied to determine the proper Q_h that matches the oscillatory level of the interference. Additionally, a range should be constrained for the investigation of the optimal Q_h. Thus, the optimal Q_h can be found as

$$Q_h = \underset{Q_h}{\arg\min}\, SI(\hat{\mathbf{y}}_l) \;\; s.t.\; Q_h \in [Q_{h\min}, Q_{h\max}] \tag{4.19}$$

The flowchart of the SI-guided parameter selection for Q_h is given in Fig. 4.13. In order for MCA to be successful, the low-oscillatory wavelets and high-oscillatory wavelets should have low mutual coherence [40], which implies that Q_h must be large enough to suppress the coherence. However, if Q_h is too high, the wavelets may not be well matched to the desired high-oscillatory component [40]. The maximum inner product of low-oscillatory wavelets and high-oscillatory wavelets is formulated as [40]

$$\rho_{\max}\left(Q_l, Q_h\right)=\sqrt{\frac{\left(Q_l+0.5\right)}{\left(Q_h+0.5\right)}} \tag{4.20}$$

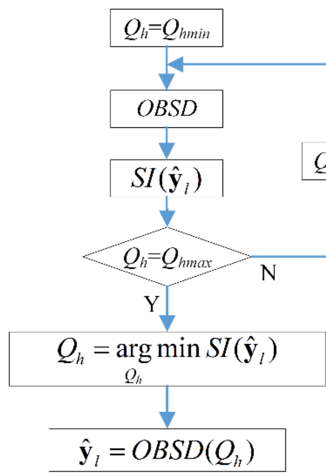

Fig. 4.13 Flowchart of SI-guided self-tuning algorithm for Q_h

With Q_l=1, ρ_{max} drops more slowly with each additional increment of Q_h [62]. When Q_h=2 and Q_h=3, ρ_{max} are 0.7746 and 0.6547 respectively, which may not be low enough to ensure the reliability of the MCA. When Q_h=4, ρ_{max} further drops to 0.5774, which is much lower than those in the cases of Q_h=2 and Q_h=3. Thus, the $Q_{h\min}$ is chosen as 4 in this chapter. When Q_h is increased to 8, ρ_{max} drops to 0.4201, which indicates ρ_{max} is decreased by 58%. With each further increment of Q_h, ρ_{max} is further decreased, however, the computational cost of the self-tuning algorithm increases with the increase of $Q_{h\max}$. Hence, the $Q_{h\max}$ is set as 8 for a desirable trade-off between a small ρ_{max} and a reasonable computational cost. The searching step of the Q_h is taken as 1. By investigating the result with minimum SI, the most suitable Q_h is found.

4.7 Simulations

4.7.1 Proposed automatic parameter selection for OBSD for bearing fault signature extraction

An automatic parameter selection method for OBSD is proposed for bearing fault signature extraction based on the index-guided parameter selection algorithms of section 4.6. The flowchart of the automatic OBSD is illustrated in Fig. 4.14. According to the analysis in section 4.5, Q_l is fixed at 1. Additionally, P_l and P_h are selected as $P_{\max}$. It has been shown that bearing fault signatures can be successfully extracted via the OBSD with r_l=r_h=5 in [69]. To further increase the overlapping extent of frequency responses of both the low-oscillatory TQWT and the high-oscillatory TQWT, r_l and r_h are chosen to be 6 in this chapter. By combining the SI-guided parameter selection algorithm for Q_h, the SI-guided parameter section algorithm for λ and the CI-guided self-tuning algorithm for μ and K, the four other parameters are then automatically chosen. The automatic parameter selection method is dedicated to searching for a pair of Q_h and λ which leads to the decomposed low-oscillatory component with minimum SI via the OBSD. As shown in Fig. 4.14, the OBSD is repeatedly applied to the original signal with different Q_h from 4 to 8. For each Q_h, the SI-guided parameter section algorithm for λ is implemented to find the proper λ. In addition, in each implementation of OBSD for finding λ, the CI-guided self-tuning algorithm for μ and K is applied. Thus, for each Q_h, the λ which results in the separated low-oscillatory component with the minimum SI is located. After applying the process to various Q_h, the Q_h and λ are selected as the pair which yields the minimum SI. At the same time, the corresponding μ and K are also selected properly.

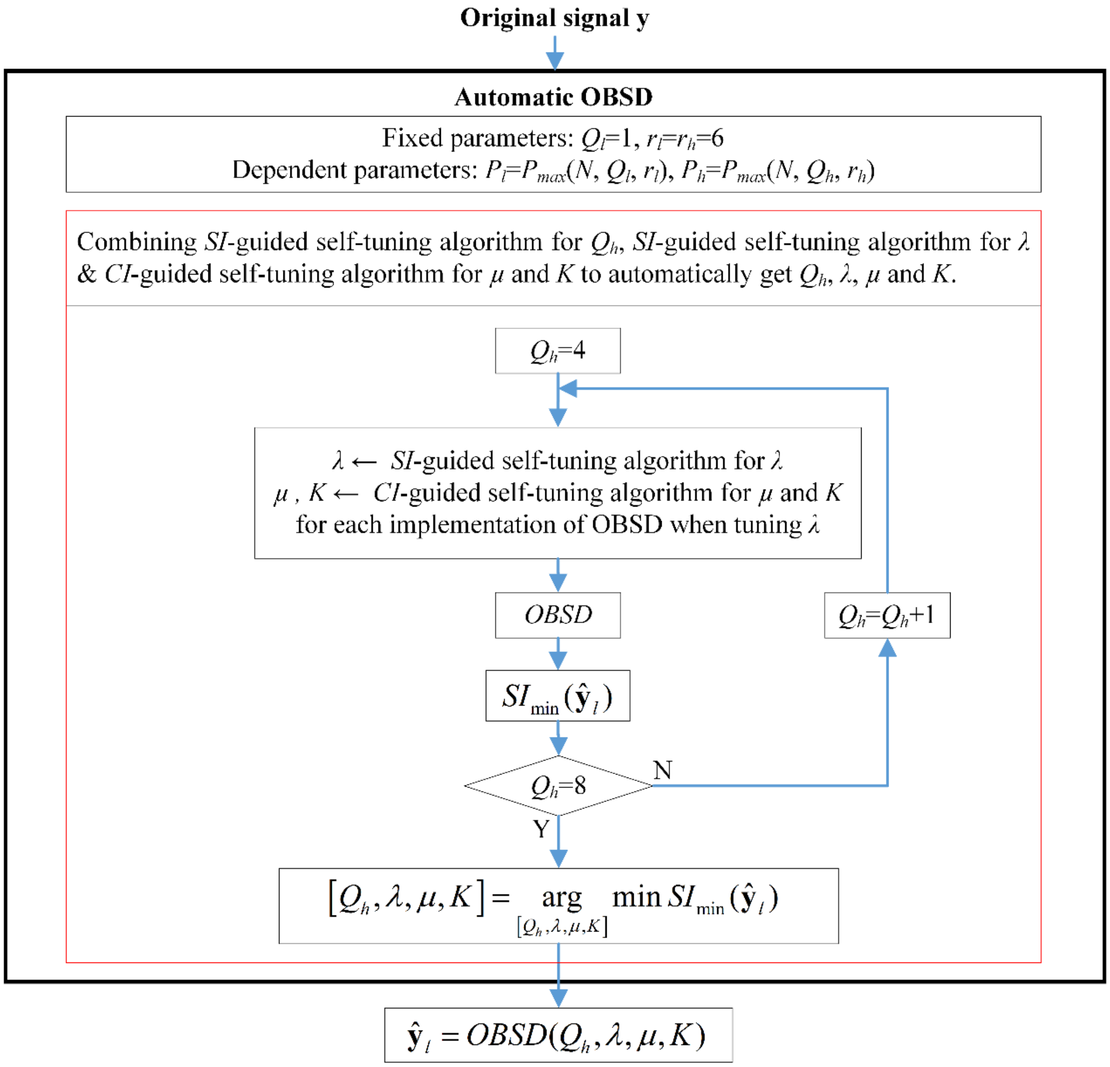

Fig. 4.14 Flowchart of the proposed Auto-OBSD for bearing fault signature extraction

An implementation of the proposed Auto-OBSD is illustrated in Fig. 4.15. The original signal (Fig. 4.15(a)) is composed of the bearing fault-induced impulses (Fig. 4.15(c)), the interference (Fig. 4.15 (e)) and noise (Fig. 4.15 (g)). The original signal is the #1 testing signal listed in Table 4.5, which is generated from Eq.(4.1), (4.3) and (4.4) with settings f_c=35Hz, f_{h1}=36Hz, f_{h2}=500Hz, SIR=-15dB, SNR=-5dB. As can be seen from the envelope spectrum of the original signal (Fig. 4.15(b)), the envelope spectrum is dominated by the frequencies of the interferences, their harmonics and sidebands. By using the proposed automatic parameter selection method, the parameters are tuned to Q_h=7, λ=0.6, μ=1 and K=124 for the OBSD. The scatter diagram of SI-λ for Q_h each obtained during the parameter tuning process is shown in Fig. 4.15(h). According to the SI-guided parameter section algorithm for λ shown in Fig. 4.11, the initial λ is set to 0.5. The searching of λ tends to decrease SI and stops at the increase of SI. For example, when Q_h=4, the SI of the low-oscillatory component obtained via the OBSD drops at λ =0.5 and λ =0.6, however, climbs at λ =0.7. Thus, for Q_h=4, λ =0.7 is taken as optimal. Similarly, the optimal values of λ are taken as 0.6, 0.7, 0.6 and 0.7 for Q_h=5, Q_h=6, Q_h=7 and Q_h=8, respectively. Considering the minimum SI at various values of Q_h which are 4,5,6,7 and 8, the Q_h is chosen as 7 according to the scatter diagram shown in Fig. 4.15(h). Hence, λ is located at 0.6 correspondingly. Also, μ and K are selected as μ=1 and K=124 according to the CI-guided self-tuning algorithm. With those parameters, the low-oscillatory component obtained via OBSD is obtained (Fig. 4.15(i)) and the corresponding envelope spectrum is shown in Fig. 4.15(j). It can be seen that the FCF are clearly shown in Fig. 4.15(j) and the interferences and their harmonics have been completely removed.

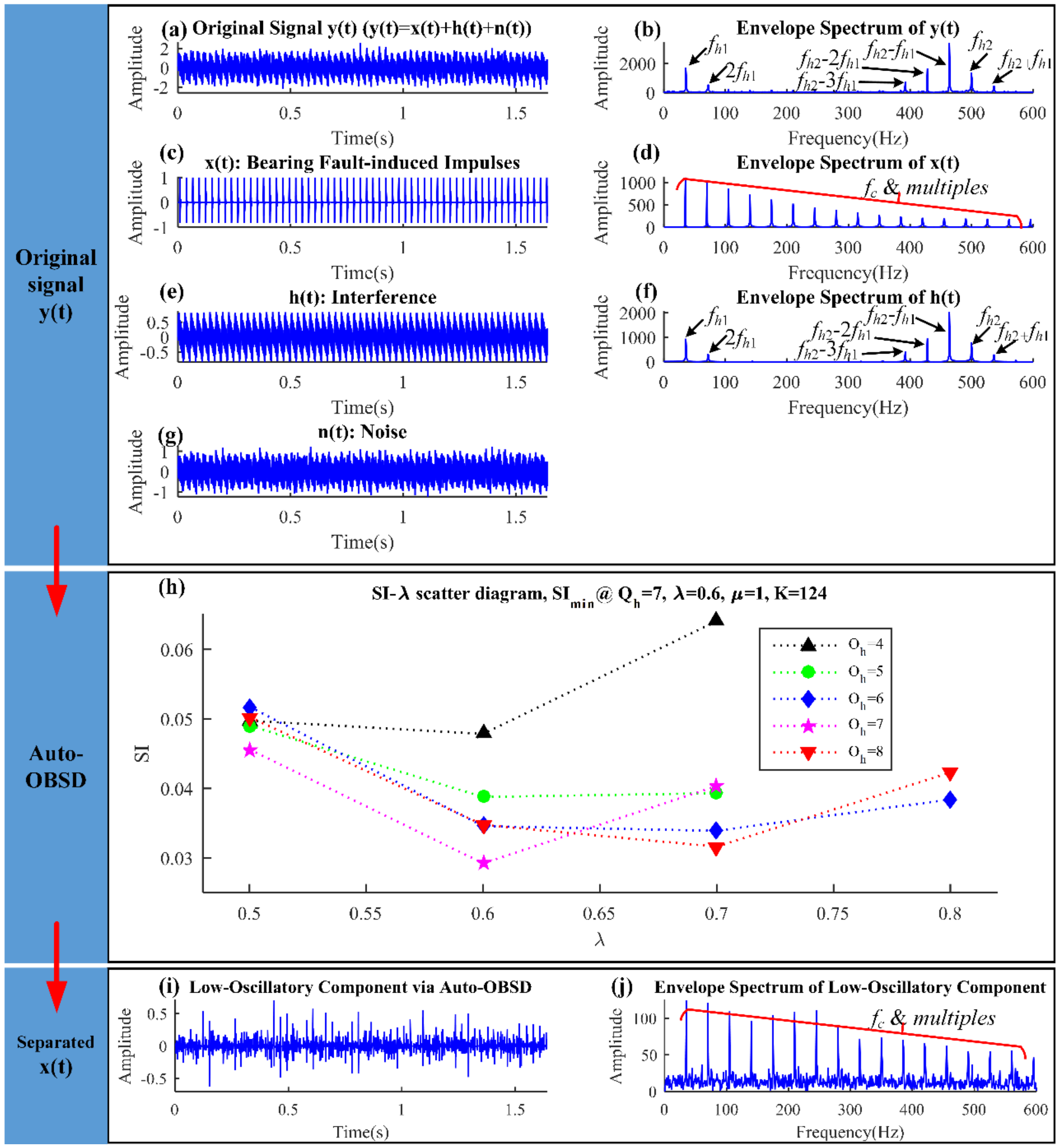

Fig. 4.15 Implementation of the proposed Auto-OBSD

4.7.2 Test signals

To validate the effectiveness of the automatic OBSD, a series of test signals are generated to test the performance of the proposed method. As mentioned in section 4.3.1, the typical observed vibration signal of a defective bearing is often modeled by Eq.(4.1), (4.3) and (4.4). It can be defined by choosing f_c, ω_r, β, τ_i, I, f_h, G, B_j, SIR and SNR. For the convenience of comparison, parameters ω_r, β, τ_i, I, G, and B_j are fixed to 2000*2π rad/s, 800, 0.01T_P, 2, 3 and [2,1, 0.5], respectively. Thus, a signal can be simply described by f_c, f_{h1}, f_{h2}, SIR and SNR in the following tests. In addition, the length of the signal N is fixed to 2^{15}=32768.

Eight different sets of signals are generated for the test. The settings of the simulated signals are listed in Table 4.5. In order to test the frequency-independent property of the proposed Auto-OBSD, two different frequencies are set for f_c (35Hz and 175Hz). Each signal is contaminated by two interference components, one with the fundamental frequency similar to f_c and the other with a relatively higher fundamental frequency. For f_c=35Hz, f_{h1}=36Hz and f_{h2}=500Hz. For f_c=175Hz, f_{h1}=177Hz and f_{h2}=1000Hz. Also, to test the performance of the proposed method on signals with different intensity of interference and noise, the SIR and SNR are set for two different values. The SIR are set as -15dB and -30dB, respectively, to represent relatively low intensity and high intensity of the interference. Similarly, SNR are set as -5dB and -12dB respectively, which refer to relatively low intensity and relatively high intensity of background noise. Hence, a total of eight different signals are constructed.

Table 4.5 Simulated signals

No.	Simulated signal	No.	Simulated signal
1	f_c=35Hz, f_{h1}=36Hz, f_{h2}=500Hz, SIR=-15dB, SNR=-5dB	5	f_c=175Hz, f_{h1}=177Hz, f_{h2}=800Hz, SIR=-15dB, SNR=-5dB
2	f_c=35Hz, f_{h1}=36Hz, f_{h2}=500Hz, SIR=-30dB, SNR=-5dB	6	f_c=175Hz, f_{h1}=177Hz, f_{h2}=800Hz, SIR=-30dB, SNR=-5dB
3	f_c=35Hz, f_{h1}=36Hz, f_{h2}=500Hz, SIR=-15dB, SNR=-12dB	7	f_c=175Hz, f_{h1}=177Hz, f_{h2}=800Hz, SIR=-15dB, SNR=-12dB
4	f_c=35Hz, f_{h1}=36Hz, f_{h2}=500Hz, SIR=-30dB, SNR=-12dB	8	f_c=175Hz, f_{h1}=177Hz, f_{h2}=800Hz, SIR=-30dB, SNR=-12dB

4.7.3 Simulation results

The results are obtained by implementing the proposed Auto-OBSD on the 8 simulated signals separately. The tuned parameters, SI values of the simulated original signals and the SI values of the decomposed low-oscillatory components are listed in Table 4.6. The simulation results for signal #1, #2, #3 and #4 are shown in Fig. 4.16 and the results for signal #5, #6, #7 and #8 are shown in Fig. 4.17, respectively.

Table 4.6 Simulation results

No.	Tuned parameter	SI value ($\mathbf{y}$)	SI value ($\hat{\mathbf{y}}_l$)	No.	Tuned parameter	SI value ($\mathbf{y}$)	SI value ($\hat{\mathbf{y}}_l$)
1	Q_h=7, λ=0.6, μ=1, K=124	0.6751	0.0292	5	Q_h=7, λ=1.3, μ=1, K=97	0.6731	0.0456
2	Q_h=6, λ=0.6, μ=1, K=258	0.6588	0.0570	6	Q_h=6, λ=1.4, μ=1, K=274	0.6648	0.1222
3	Q_h=6, λ=0.9, μ=1.5, K=89	0.6782	0.0363	7	Q_h=6, λ=1.5, μ=1, K=99	0.6778	0.1377
4	Q_h=7, λ=1.1, μ=1, K=220	0.6633	0.0553	8	Q_h=8, λ=1.5, μ=1, K=257	0.6651	0.0612

It can be seen from Table 4.6 that the SI value of the decomposed low-oscillatory component is much smaller than the SI value of the original signal in each case. It reveals that the smoothness of the results is greatly reduced, which implies that the results are much closer to the target impulsive bearing fault-induced signals. As shown in Fig. 4.16 and Fig. 4.17, the decomposed low-oscillatory components are all more impulsive than the corresponding original signals. Moreover, the envelope spectra of the low-oscillatory components show more clear peaks at the FCF and its harmonics than the envelope spectra of original signals. It can be concluded that the minimum SI criterion can indeed guides proper parameter selection for the OBSD in the application of bearing fault signature extraction. However, the values of the SI do not necessarily represent the quality of the results in different cases. For example, the SI value of the decomposed low-oscillatory component of signal #2 is 0.0570 which is greater than that of the signal #3, which is 0.0363. However, the envelope spectrum of the result of signal #2 shows more clear peaks at the FCF and its harmonics, which implies that the result of signal #2 is better than that of signal #3. Hence, the minimization of SI can guide the result to be close to the desired signal for a single case of bearing fault signature extraction by using the OBSD. However, the SI may not be used to judge the quality of the result for different cases of original signals.

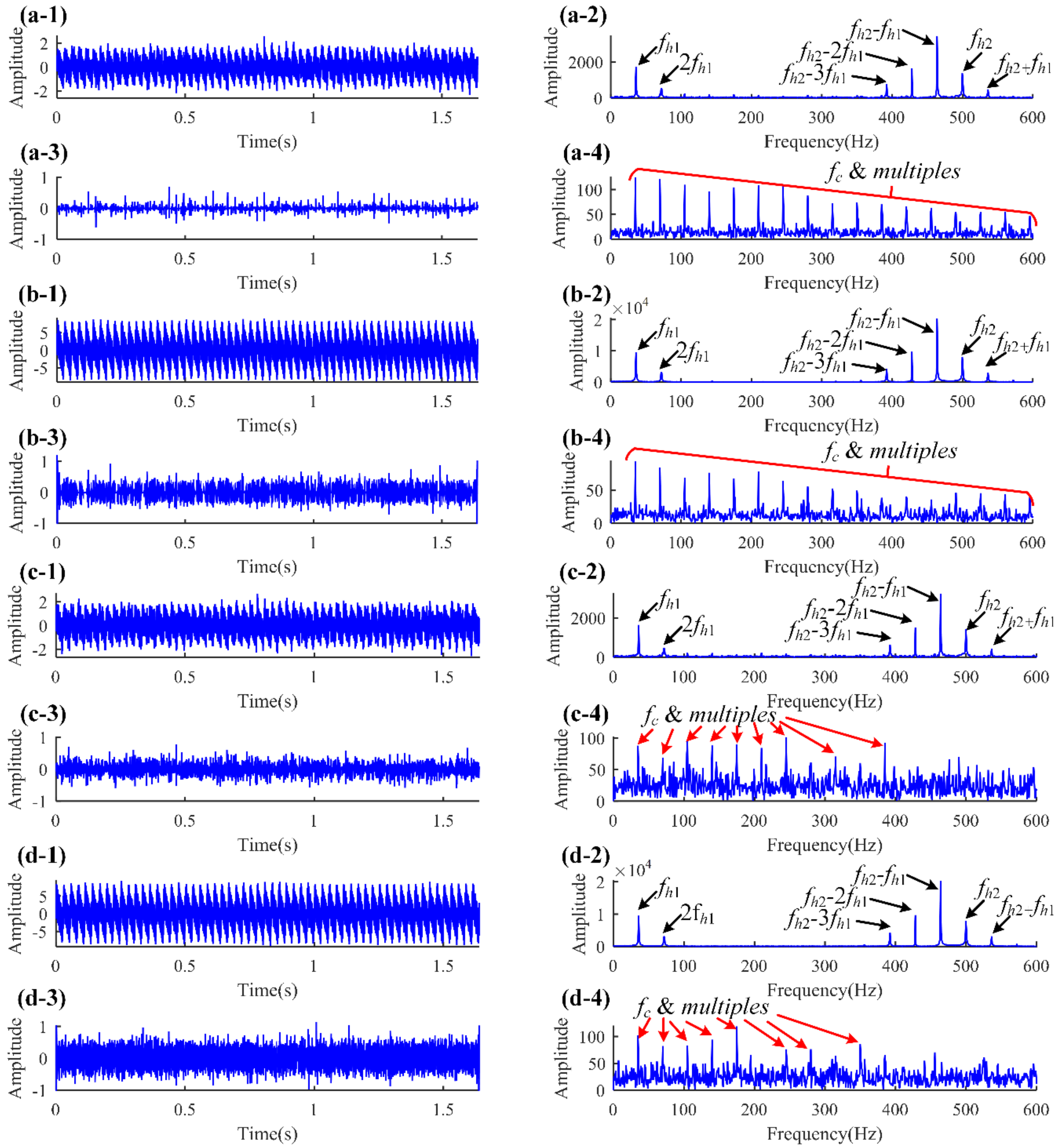

Fig. 4.16 Simulation results: (a-1) test signal #1 in Table 4.5, (a-2) envelope spectrum of (a-1), (a-3) low-oscillatory component of (a-1) via OBSD with self-tuned parameters, (a-4) envelope spectrum of (a-3); (b-1) test signal #2, (b-2) envelope spectrum of (b-1), (b-3) low-oscillatory component of (b-1), (b-4) envelope spectrum of (b-3); (c-1) test signal #3, (c-2) envelope spectrum of (c-1), (c-3) low-oscillatory component of (c-1), (c-4) envelope spectrum of (c-3); (d-1) test signal #4, (d-2) envelope spectrum of (d-1), (d-3) low-oscillatory component of (d-1) and, (d-4) envelope spectrum of (d-3)

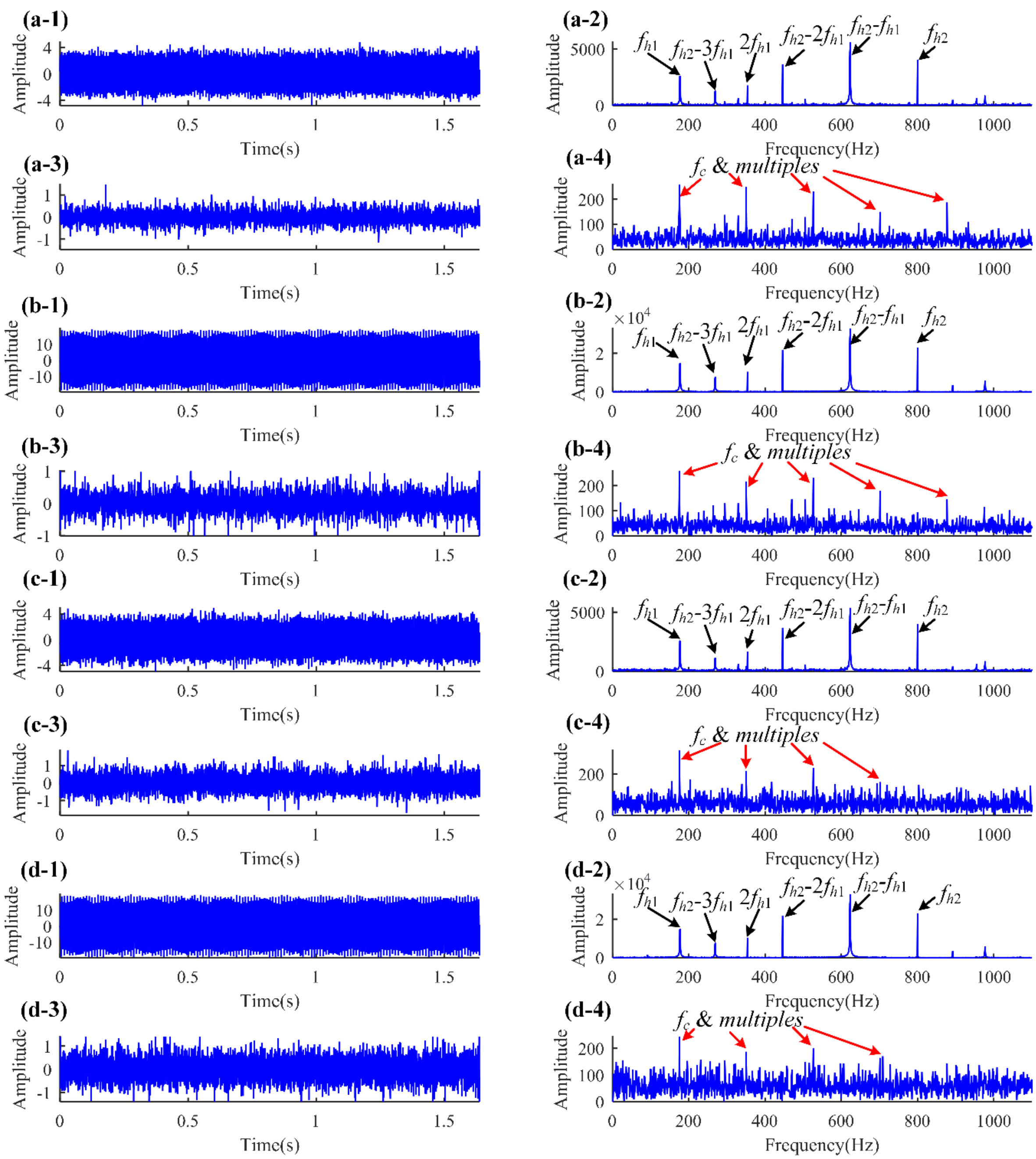

Fig. 4.17 Simulation results: (a-1) test signal #5 in Table 5, (a-2) envelope spectrum of (a-1), (a-3) low-oscillatory component of (a-1) via OBSD with self-tuned parameters, (a-4) envelope spectrum of (a-3); (b-1) test signal #6, (b-2) envelope spectrum of (b-1), (b-3) low-oscillatory component of (b-1), (b-4) envelope spectrum of (b-3); (c-1) test signal #7, (c-2) envelope spectrum of (c-1), (c-3) low-oscillatory component of (c-1), (c-4) envelope spectrum of (c-3); (d-1) test signal #8, (d-2) envelope spectrum of (d-1), (d-3) low-oscillatory component of (d-1) and, (d-4) envelope spectrum of (d-3)

The tuned Q_h for simulated signals listed in Table 4.6 are not the same. Theoretically, the optimal Q_h for each test signal should be the same since the oscillatory behavior of the interference of each signal is the same. The reason is that the interval of λ for parameter tuning is 0.1, which may not be precise enough to lead an optimal result for each Q_h. Therefore, the optimal Q_h may not be selected via the automatic parameter selection method. However, it is not necessary to select the optimal Q_h to obtain a good result. As can be seen in Fig. 4.17, the results of signal #1 and signal #2 are both good enough to detect the bearing fault, even though the Q_h are selected as different values.

The proposed automatic parameter selection method works for signals at different settings of f_c, f_{h1} and f_{h2}. The original signals shown in Fig. 4.16 and in Fig. 4.17 are composed of bearing fault-induced impulses and interferences at different frequencies. Comparing the signals with the same SNR and SIR, the characteristics of the results are similar. The signal #1 shown in Fig. 4.16(a-1) and signal #5 shown in Fig. 4.17(a-1) are signals with different f_c, f_{h1} and f_{h2}, the envelope spectra shown in Fig. 4.16(a-2) and in Fig. 4.17(b-2) are both dominated by the frequencies of the interferences, their harmonics and sidebands. However, from the envelope spectra of the decomposed low-oscillatory component, the FCF and its harmonics are clearly shown in Fig. 4.16(a-4) and in Fig. 4.17(b-4). In addition, the peaks at the frequencies of the interferences, their harmonics and sidebands have all disappeared.

The simulation results have also demonstrated ability to separate the bearing fault features from a signal with strong interference. The original signals shown in Fig. 4.16 are composed of bearing fault-induced signals, strong interferences with SIR=-30dB and relative weak noise with SNR=-5dB. As shown, the fault frequencies are all clearly discernible in the results. However, the performance is affected by the strength of noise. As can be seen in Fig. 4.16 and Fig. 4.17, the results of the signals with relatively low intensity of noise with SNR=-5dB are satisfactory. The FCFs and their harmonics are clearly shown in the envelope spectra. However, the results of those with relatively high level noise with SNR=-12dB are not as good as the former ones, though the bearing fault can still be identified from the result.

The computational cost of the Auto-OBSD depends on the computational cost of OBSD and the input signal. According to [62], the computational cost of the TQWT is $O(rN\log_2 N)$. Then, the computational cost of the OBSD is $O(rKN\log_2 N)$. Therefore, the computational cost of the Auto-OBSD is $O\left(r\sum_{i=1}^{n}K_i N\log_2 N\right)$ where K_i is the K of the i^{th} repeated OBSD, and n is the number of times the OBSD calculated throughout the parameter tuning process. The n depends on the signals. In our simulations, n varied between 15 and 50. Thus, the computational cost of the Auto-OBSD depends linearly on the computational cost of the OBSD, which is relatively low as can be expected.

4.8 Experimental Evaluations

To further validate the effectiveness of the proposed parameter selection method, it is applied to experimental data. In this chapter, two types of bearing faults are considered: defects on the outer race and defects on the inner race.

4.8.1 Experimental setup

The experiments are performed on a SpectraQuest machinery fault simulator (MFS-PK5M). The experimental set-up is shown in Fig. 4.18. The shaft is supported by two ER16K ball bearings, the left one is healthy and the right one is faulty. The rotation of the shaft is transmitted to the gearbox by two belt sheaves. The diameter ratio of the two sheaves is 1:2.6 (the small one is connected to the shaft and the large one is connected to the gearbox). The vibration signal is collected by an accelerometer (ICP accelerometer, Model 623C01) located on the base of the test rig, and the data is sampled by an NI data acquisition board (NI USB-6212 BNC) and LABVIEW. With this configuration, the collected signal contains both the bearing fault-induced impulses and the gear meshing signal (interference). According to the structure parameters of the bearings given in Table 4.7, the fault characteristic frequency (FCF) of the bearing outer race, i.e. Ball-Pass Frequency of Outer-race (BPFO), is equal to the product of the FCF coefficient (3.57) and the shaft rotational frequency f_r, i.e. f_c=3.57f_r. Similarly, the FCF of the bearing inner race, also named as Ball-Pass Frequency of Inner race (BPFI), is 5.43f_r. In addition, the gear meshing frequency is calculated as the shaft rotational frequency divided by the diameter ratio of the sheaves and then multiplied by the number of the pinion teeth, i.e. $f_{meshing}$=(18/2.6)f_r =6.92f_r.

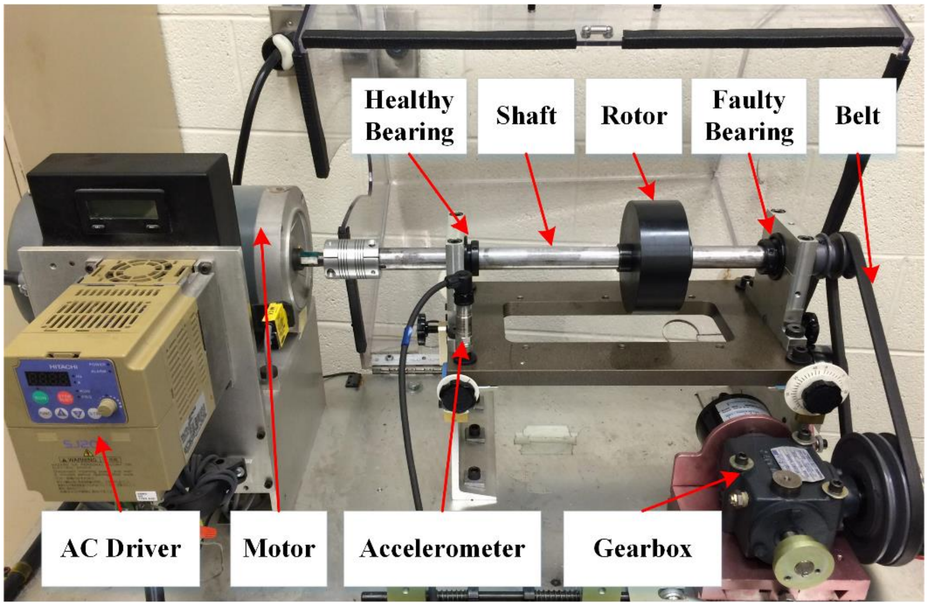

Fig. 4.18 Experimental setup

Table 4.7 Parameters of bearings and gears

Bearing type	Pitch diameter	Ball diameter	Number of balls	BPFO	BPFI
ER16K	38.52mm	7.94mm	9	$3.57f_r$	$5.43f_r$
Diameter ratio of sheaves		Number of teeth		Gear meshing frequency	
1:2.6		18		$6.92f_r$	

4.8.2 Bearing outer race fault detection

For the outer race fault case, a bearing with a local defect on the outer race is mounted as the faulty bearing on the test rig shown in Fig. 4.18. The shaft rotational frequency was set at 15Hz, the signal sampling rate was set at 20kHz and the number of samples was 32748 (signal length was 1.6374s). Given this data, the bearing fault characteristic frequency, i.e. BPFO, and the gear meshing frequency are calculated as 53.55Hz and 103.8Hz, respectively. The measured signal is shown in Fig. 4.19(a) and the corresponding envelope spectrum is shown in Fig. 4.19(b). It can be seen that the gear meshing frequency shows a clear peak in the envelope spectrum while the bearing fault characteristic frequency is totally submerged. Obviously, the bearing fault cannot be detected by the envelope spectrum of the original signal.

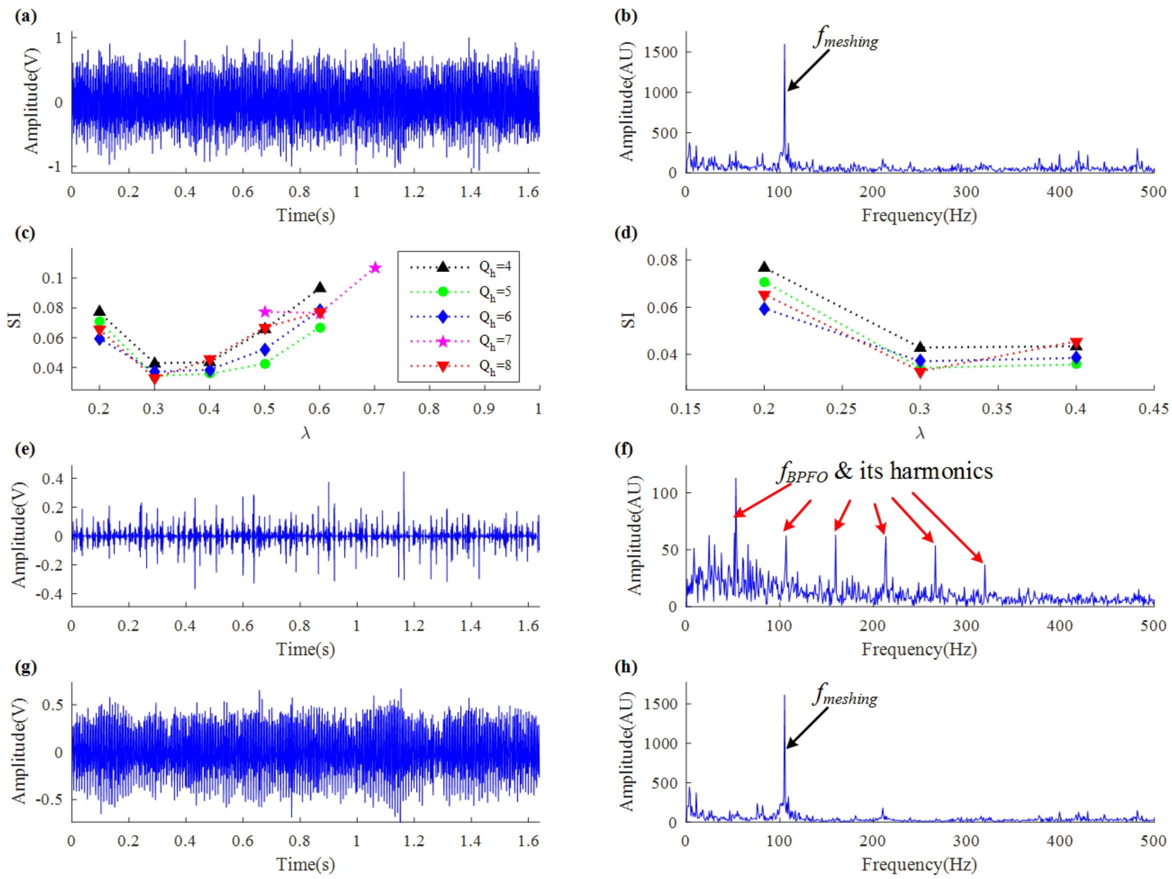

Fig. 4.19 Implementation of the Auto-OBSD for bearing fault signature extraction on bearing with outer race fault: (a) original signal, (b) envelope spectrum of (a), (c) SI-λ scatter diagram by using the proposed automatic parameter selection method, (d) local zoom-in version of (c), (e) decomposed low-oscillatory component via the OBSD with the self-tuned parameters, (f) envelope spectrum of (e), (g) decomposed high-oscillatory component, and (h) envelope spectrum of (g)

The SI-λ scatter diagram is shown in Fig. 4.19(c) by applying the proposed automatic parameter selection method. At a range of Q_h=[4,8], the SI-guided parameter section algorithm for λ and CI-guided self-tuning μ and K algorithm are applied. From the zoomed-in version of the SI-λ scatter diagram, it can be seen that the SI value of the low-oscillatory component obtained via the OBSD reaches the minimum with Q_h=8, λ=0.3. The corresponding μ and K are μ=1 and K=111.

The decomposed low-oscillatory component and its envelope spectrum are shown in Fig. 4.19(e) and Fig. 4.19(f), respectively. It can be seen that the signal becomes more impulsive than the collected original signal. Additionally, the FCF and its harmonics show very clear peaks in the envelope spectrum. Moreover, the peak of the gear meshing frequency is removed from the envelope spectrum. It reveals that the interference, i.e. gear meshing signal, is totally separated from the original signal.

4.8.3 Bearing inner race fault detection

For the inner race fault case, faulty bearing on the test rig shown in Fig. 4.18 is replaced by the bearing with a local defect on the inner race. The shaft rotational frequency was set at 17Hz, the signal sampling rate was also set at 20kHz and the number of samples was 32748 (signal length was 1.6374s). With this setting, the bearing fault characteristic frequency, i.e. BPFI, and the gear meshing frequency are calculated as 92.31Hz and 117.64Hz, respectively. The measured signal is shown in Fig. 4.20(a) and the corresponding envelope spectrum is shown in Fig. 4.20(b). Similar to the outer race case, the envelope spectrum is dominated by the gear meshing frequency. Therefore, the bearing fault cannot be detected by the envelope spectrum of the original signal.

The SI-λ scatter diagram obtained by the proposed Auto-OBSD is shown in Fig. 4.20(c). In a range of Q_h=[4,8], the SI-guided parameter section algorithm for λ and CI-guided self-tuning μ and K algorithm are applied. From the zoomed-in version of the SI-λ scatter diagram, it can be seen that the SI value of the low-oscillatory component obtained via the OBSD reaches the minimum with Q_h=6, λ=0.2. The corresponding μ and K are μ=1 and K=107.

The decomposed low-oscillatory component and its envelope spectrum are shown in Fig. 4.20(e) and Fig. 4.20(f), respectively. It can be seen that the impulses are more visible in the decomposed low-oscillatory component than in the original signal. Additionally, the FCF and its harmonics show very clear peaks in the envelope spectrum. For the inner race case, the sidebands related to the shaft rotating frequency are also shown in the envelope spectrum.

Furthermore, the peak of the gear meshing frequency is removed from the envelope spectrum. It indicates that the interference, i.e. gear meshing signal, is fully separated from the original signal.

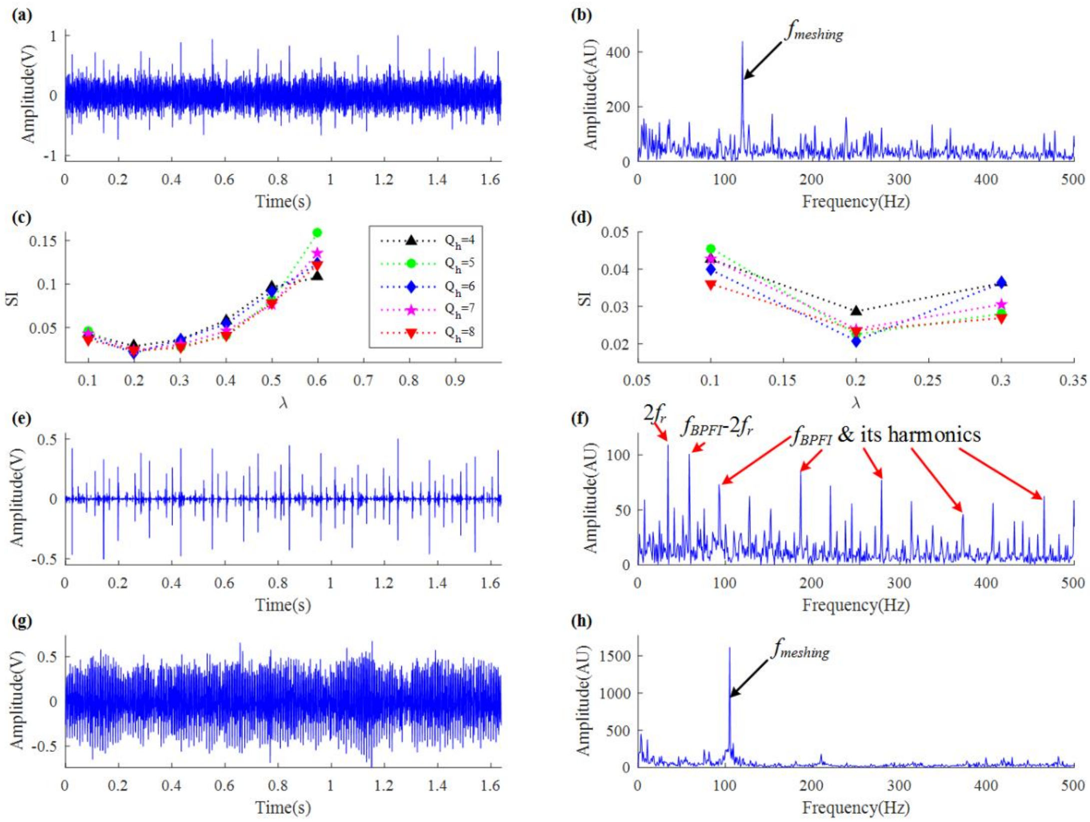

Fig. 4.20 Implementation of the Auto-OBSD for bearing fault signature extraction on bearing with inner race fault: (a) original signal, (b) envelope spectrum of (a), (c) SI-λ scatter diagram by using the proposed automatic parameter selection method, (d) local zoom-in version of (c), (e) decomposed low-oscillatory component via the OBSD with the self-tuned parameters, (f) envelope spectrum of (e), (g) decomposed high-oscillatory component, and (h) envelope spectrum of (g)

4.8.4 Discussions

From the experimental validation, it can be seen that the proposed parameter selection method can be used to automatically tune the parameters for the OBSD for bearing fault signature extraction. The results indicate that the proposed method works well for bearings with an outer race fault or with an inner race fault.

It is also found that the automatically selected Q_h for outer race case and inner race case are different. However, both results are sufficient for bearing fault detection. This confirms the analysis of the simulation results in section 4.7.3.

4.9 Conclusions

In this chapter, the effects of the parameter selection on the OBSD performance in bearing fault signature extraction were analyzed. It was found that the main parameters needed to be tuned are high Q-factor Q_h, regularization parameter λ, penalty factor μ and the number of iterations K. Based on the analysis, the CI was proposed to guide the self-tuning of μ and K. Additionally, the SI was also employed to guide the automatic selection of Q_h and λ. With the three index guided self-tuning algorithms, an automatic parameter selection method was developed for the OBSD for bearing fault signature extraction. By applying the proposed method to a series of simulated signals, it was found that the method performed well even when the interference frequency is almost identical to the fault characteristic frequency and when signal-to-interference ratio is very low. Furthermore, the method was implemented on experimental data collected from a bearing with outer race fault and inner race fault, respectively. The results indicated that the method was able to extract the bearing fault signature from the signal contaminated by gear meshing signals for both the outer race case and inner race case. It should also be noted that the above results were achieved without any pre-denoising steps.

Chapter 5 Bearing fault diagnosis under unknown time-varying rotational speed conditions via multiple time-frequency curve extraction

This chapter addresses objective 3. A new method based on the time-frequency analysis was proposed for bearing fault diagnosis under time-varying speed conditions without the presence of interference signals. The proposed method is resampling-free and tachometer-free.

The content of this chapter has been published in the *Journal of Sound and Vibration*.

Huan Huang, Natalie Baddour, and Ming Liang. "Bearing fault diagnosis under unknown time-varying rotational speed conditions via multiple time-frequency curve extraction." *Journal of Sound and Vibration* 414 (2018): 43-60.

5.1 Abstract

Under normal operating conditions, bearings often run under time-varying rotational speed conditions. Under such circumstances, the bearing vibrational signal is non-stationary, which renders ineffective the techniques used for bearing fault diagnosis under constant running conditions. One of the conventional methods of bearing fault diagnosis under time-varying speed conditions is resampling the non-stationary signal to a stationary signal via order tracking with the measured variable speed. With the resampled signal, the methods available for constant condition cases are thus applicable. However, the accuracy of the order tracking is often inadequate and the time-varying speed is sometimes not measurable. Thus, resampling-free methods are of interest for bearing fault diagnosis under time-varying rotational speed for use without tachometers. With the development of time-frequency analysis, the time-varying fault character manifests as curves in the time-frequency domain. By extracting the Instantaneous Fault Characteristic Frequency (IFCF) from the Time-Frequency Representation (TFR) and converting the IFCF, its harmonics, and the Instantaneous Shaft Rotational Frequency (ISRF) into straight lines, the bearing fault can be detected and diagnosed without resampling. However, so far, the extraction of the IFCF for bearing fault diagnosis is mostly based on the assumption that at each moment the IFCF has the highest amplitude in the TFR, which is not always true. Hence, a more reliable T-F curve extraction approach should be investigated. Moreover, if the T-F curves including the IFCF, its harmonic, and the ISRF can be all extracted from the TFR directly, no extra processing is needed for fault diagnosis. Therefore, this chapter proposes an algorithm for multiple T-F curve extraction from the TFR based on a fast path optimization which is more reliable for T-F curve extraction. Then, a new procedure for bearing fault diagnosis under unknown time-varying speed conditions is developed based on the proposed algorithm and a new fault diagnosis strategy. The average curve-to-curve ratios are utilized to describe the relationship of the extracted curves and fault diagnosis can then be achieved by comparing the ratios to the fault characteristic coefficients. The effectiveness of the proposed method is validated by simulated and experimental signals.

5.2 Introduction

Vibration signal analysis is one of the most widely used techniques for bearing fault diagnosis [17]. Once a local bearing defect is developed, it induces an impulse excitation to the vibration system and generates impulsive responses in the vibration signal which amplitude-modulate the carrier frequencies, i.e., the high resonance frequencies [9]. In addition, the impulsive responses appear at a certain frequency called the Fault Characteristic Frequency (FCF). For each type of fault, its FCF is given as the product of the shaft rotational frequency and a specific coefficient related to the dimensions of the bearing [9]. Therefore, by observing the FCF and its harmonics in the frequency spectrum of the demodulated vibration signal, bearing faults can be detected and diagnosed. However, in real applications, bearings often operate under time-varying rotational speed conditions. In such cases, the signal is non-stationary and the FCF is no longer constant. Thus, bearing faults cannot be diagnosed in the frequency spectrum directly.

One of the conventional methods for use with vibration signals for time-varying speed cases is signal resampling via order tracking. The essence of order tracking is the resampling of the vibration signal at a constant angle interval rather than constant time interval [34]. After resampling, the time-varying FCF present in the signal sampled at constant time intervals is converted into a constant fault characteristic order (multiple of the reference shaft speed) in the resampled signal [31]. Subsequently, bearing fault diagnosis can be conducted in the order domain. In addition, tachometer-free order-tracking methods have been proposed for bearing fault diagnosis under time-varying speed conditions ([32],[93]), since the shaft rotational speed is not always measurable due to the lack of measuring devices and/or the infeasibility of installation of such devices. However, the accuracy of order tracking is impacted by factors such as the precision of the obtained rotational speed and the method of interpolation [34]. Moreover, the order tracking process is affected by the transmission path phase [94]. Thus, resampling free methods are of particular interests for bearing fault diagnosis under time-varying speed conditions.

Time-Frequency analysis enables resampling-free methods for machinery fault diagnosis under time-varying speed conditions, since the time-varying fault characters appear as “ridges” in the well processed Time-Frequency Representation (TFR) [70]. Various time-frequency analysis methods have been developed to obtain the TFR, among which the Short-Time Fourier Transform (STFT) and wavelet transform are the most popular [70]. Furthermore, to improve the readability of the TFR, synchrosqueezed wavelet transforms have been developed by Daubechies et al [95]. However, the synchrosqueezing method could be handicapped by time-dimension diffusions of the wavelet coefficients. To deal with diffusions in both time and frequency dimensions, a generalized synchrosqueezing transform approach was proposed by Li et al [96].

Methods based on Generalized Demodulation (GD) have been investigated for using the TFR for gearbox fault diagnosis under time-varying speed conditions by Feng et al ([97]–[99]). Generalized demodulation can be utilized to convert the arbitrary instantaneous frequency trajectories into constant frequency lines in the time-frequency plane. By iteratively applying generalized demodulation, multiple instantaneous frequency trajectories can be converted into constant frequency lines. A similar strategy has been investigated and applied to bearing fault diagnosis under time-varying speed conditions by Shi et al [81]. However, to apply generalized demodulation, the instantaneous frequency of the target signal component, i.e. the Instantaneous Fault Characteristic Frequency (IFCF), has to be estimated. Theoretically, the IFCF and its harmonics manifest themselves as ridges in the TFR and ideally the IFCF shows a more significant ridge than its harmonics. Thus, in [81], the IFCF is estimated as the frequency with the maximum amplitude at each moment in the TFR obtained via the STFT. However, the peak of the FCF is not always the highest peak in the frequency spectrum. Thus, a multiple amplitude superposition algorithm was proposed to extract the IFCF by Wang et al [32]. However, extracting the Time-Frequency (T-F) curve by relying only on the amplitude of the peaks may result in the extracted curve not being consistent with the curve itself, for example the extracted curves contains unexpected frequency jumps. The results can be worse if the signal is noisy. Therefore, more reliable methods for T-F curve extraction from the TFR need to be investigated. Moreover, if multiple T-F curves can be directly extracted from the TFR, then it is not necessary to apply generalized demodulation to convert the instantaneous frequency trajectories into constant frequency lines. By doing so, the process of bearing fault diagnosis can be simplified and the computational efficiency can be improved accordingly.

T-F curve extraction (also called Instantaneous Frequency (IF) extraction or ridge extraction) from the TFR has been investigated by many researchers ([100]–[103]). The TFR can be regarded as an image, thus, image processing technique can be applied to locate the ridges in the TFR. An algorithm composed of steps of image binary, dilatation, and skeleton operation was proposed by Borda et al to estimate the IF [100]. The IF is estimated as the result of the skeleton obtained. Another ridge extraction method based on the image processing technique of active contours was proposed by Terrin et al for the analysis of uterine electromyogram [101]. However, the connectivity of the extracted IF or ridge is not ensured via these image processing techniques since the result is a skeleton or contour composed of pixels without direction. A ridge extraction method by use of a dynamic programming algorithm was proposed by Liebling et al for spectral interferometry imaging [102]. The extracted ridge is a path composed of points with directions via this method. Following a similar idea, a T-F curve extraction approach, called fast path optimization, was proposed by Iatsenko et al [103]. The T-F curve is extracted by solving a chosen optimal function composed of three terms, including one related to the amplitude of the peaks in the TFR and the other two related to the frequency and the changes in the frequency of the peaks, respectively. The latter two terms can effectively prevent the unexpected frequency jumps which lead to the extracted T-F curve more accurately representing the ridge in the TFR. Therefore, with this approach, the T-F curve is extracted from the TFR more reliably. However, fast path optimization has not been used for T-F curve extraction from the TFR of bearing vibration signals. Moreover, bearing fault diagnosis requires multiple T-F curves to be extracted, including those corresponding to the IFCF and its harmonics, and the T-F curve associated with the Instantaneous Shaft Rotational Frequency (ISRF).

In this chapter, an algorithm for multiple time-frequency curve extraction for the purpose of bearing fault diagnostics is proposed based on the fast path optimization approach. The rest of the chapter is structured as follows. In section 5.3, the fast path optimization approach is introduced and a multiple T-F curve extraction algorithm is proposed based on the fast path optimization. In section 5.4, a procedure of bearing fault diagnosis under time-varying speed conditions is developed based on the multiple T-F curve extraction and a fault diagnosis strategy is designed accordingly. Then, in section 5.5, simulations are conducted to illustrate the effectiveness of the proposed method. Signals of linearly varying and non-linearly varying rotational frequencies are used. Additionally, the signal of linearly varying rotating speed is categorized as IFCF dominant or not dominant in the TFR to cover more types of situations. Following the simulations, in section 5.6, experimental data are used to further test the effectiveness of the proposed method. Vibration signals collected from a bearing with an outer race defect, a bearing with an inner race defect, and a healthy bearing are utilized for the validation. Finally, conclusions are drawn based on the results of simulations and the experimental validation.

5.3 Multiple time-frequency curve extraction

Extracting the time-frequency curves from the TFR can be accomplished by extracting the ridges in the TFR and finding the corresponding time-frequency curves where the ridges are located. One of the simplest approaches for ridge extraction from the TFR is extracting the ridge point-by-point which is called a “one-step” approach [103]. The procedure of the “one-step” is locating an initial ridge point first and then searching for the ridge points forwards and backwards following specific optimal functions. This approach is simple and straightforward. However, it is not reliable since one wrongly picked point may totally change the result. Thus, a more reliable approach is proposed based on fast path optimization for ridge extraction [103]. Additionally, for multiple ridge curve extraction, a common way is to find the curve associated with the dominant component, which can then be reconstructed and subtracted from the signal. Subsequently, the other curves can then be extracted by repeating the procedure. In this chapter, an algorithm for multiple time-frequency curve extraction is proposed based on the fast path optimization approach without signal reconstruction.

5.3.1 Time-frequency curve extraction via fast path optimization

Time-frequency curve extraction from the TFR can be performed by finding the ridges in the TFR and taking the corresponding frequencies along the time span as the time-frequency curve. Thus, the ridge extraction is critical for time-frequency extraction. In this chapter, the STFT is applied to obtain the TFR, the TFR of signal $x(t)$ is

$$X(\tau, f) = \int_{-\infty}^{+\infty} x(t) w(t-\tau) \mathrm{e}^{-\mathrm{j}2\pi ft} \mathrm{d}t \tag{5.1}$$

where $w()$ is the window function, τ is the time variable, and f is the frequency variable. For simplicity, τ and f are regarded as continuous variables in Eq.(5.1). However, in practice, both τ and f are discretized. In the following, we assume that the length of the TFR in time span is N and the length of the TFR in frequency span is N_f. Generally, ridge extraction is conducted by choosing a start point and locating the connected ridges forwards or backwards point-by-point along the time span following a certain optimization function. The main drawback of this approach as mentioned earlier is that a wrongly selected point may completely change the extracted ridge. Thus, a more accurate approach to extract the curve is via optimizing the whole profile of the curve, which is called path optimization. However, the computational cost will be significantly increased if the optimization is performed over all possible peaks directly. To take advantage of the path optimization approach and to reduce the high computational cost, a fast path optimization is proposed in [103].

The fast path optimization is dependent on only a finite number of the past points, rather than all peaks [103]. To demonstrate the approach, the number of peaks at time τ_n (n=1,2,…,N) in the TFR is denoted as $N_p(\tau_n)$, the amplitude of mth peak at τ_n is denoted as $Q_m(\tau_n)$, and the corresponding frequency of the mth peak at τ_n is denoted as $\nu_m(\tau_n)$. All peaks at τ_n can be determined via

$$\nu_m(\tau_n) = f \quad s.t. \begin{cases} \mathrm{d}[X(\tau_n, f)]/\mathrm{d}f = 0 \\ \mathrm{d}^2[X(\tau_n, f)]/\mathrm{d}f^2 < 0 \end{cases}, m = 1, 2, ..., N_p \tag{5.2}$$

All the peaks together generate a peak map, Fig. 5.1(a). The circles denote the peaks in the TFR. To extract a time-frequency curve from the TFR, one needs to determine which peak should be extracted as the ridge at each moment. Thus, for a TFR that has a time span $[\tau_{1,...,}\tau_N]$, the path optimization can be described as [103]

$$\left\{m_c(\tau_1), ..., m_c(\tau_N)\right\} = \underset{\{m_1,...,m_N\}}{\arg\max} \sum_{n=1}^{N} F\left[\tau_n, Q_{m_n}(\tau_n), \nu_{m_n}(\tau_n), \left\{\nu_{m_1}(\tau_1), ..., \nu_{m_N}(\tau_N)\right\}\right] \tag{5.3}$$

where $m_c(\tau_n)$ determines the peak to be extracted as the ridge at τ_n, F[] is the chosen support function for the optimization and $\left\{m_1, ..., m_N\right\}$ refers to a sequence of the peak numbers along the time span. The path optimization is performed over all peaks at the whole time span. The computational complexity is on the order of $O\left(N_p(\tau_1)\cdot N_p(\tau_2)\cdots N_p(\tau_N)\right)$ which can be up to $O\left({M_p}^N\right)$ with $M_p = Max\left(N_p(\tau_n)\right)$. Additionally, all peaks in the TFR have to be determined prior to the extraction. Therefore, the computational complexity is on the order of $O\left(N_f N + {M_p}^N\right)$ in total. It can be seen that the computational cost is extremely high if the optimization is directly applied to all possible peaks. Thus, a fast path optimization is proposed to be only dependent on a finite number of the previous points with the support function [103]

$$F[\ ] = \begin{cases} \log Q_m(\tau_n) & n=1 \\ \log Q_m(\tau_n) + w_2\left(\nu_m(\tau_n), \mathrm{m}[f_d], \mathrm{IQR}[f_d], \lambda_2\right) & n=2 \\ \log Q_m(\tau_n) + w_2\left(\nu_m(\tau_n), \mathrm{m}[f_d], \mathrm{IQR}[f_d], \lambda_2\right) & \\ +w_1\left(\nu_m(\tau_n) - f_d(\tau_{n-1}), \mathrm{m}[\Delta f_d], \mathrm{IQR}[\Delta f_d], \lambda_1\right) & n \geq 3 \end{cases} \tag{5.4}$$

where

$$w_1\left(\nu_m(\tau_n) - f_d(\tau_{n-1}), m[\Delta f_d], IQR[\Delta f_d], \lambda_1\right) = -\lambda_1 \left| \frac{\nu_{m_n}(\tau_n) - f_d(\tau_{n-1}) - m[\Delta f_d]}{IQR[\Delta f_d]} \right| \tag{5.5}$$

$$w_2\left(\nu_m(\tau_n), m[f_d], IQR[f_d], \lambda_2\right) = -\lambda_2 \left| \frac{\nu_m(\tau_n) - m[f_d]}{IQR[f_d]} \right| \tag{5.6}$$

$$\mathrm{m}[\ ] = \mathrm{perc}_{0.5}[\ ], \quad \mathrm{IQR}[\ ] = \mathrm{perc}_{0.75}[\ ] - \mathrm{perc}_{0.25}[\ ] \tag{5.7}$$

where $f_d(\tau_{n-1})$ is the frequency of the candidate ridge point at τ_{n-1}, f_d are the frequencies of a series of candidate ridge points in history $[\tau_1,...,\tau_{n-1}]$, Δf_d is the derivative of f_d, $\mathrm{m}\left[f(t)\right]$ is the median of $f(t)$, IQR is the interquartile range, defined in Eq.(5.7), where $\mathrm{perc}_p[f(t)]$ denotes the pth quantile of $f(t)$, and λ_1 and λ_2 are penalty factors which can be taken as 1 according to the analysis in [103]. Weight functions $w_1()$ and $w_2()$ are employed to suppress the atypical variations of the ridge frequency's value and derivative, respectively. With the chosen function given in Eq.(5.4), the frequency jumps can be prevented in the extracted T-F curve.

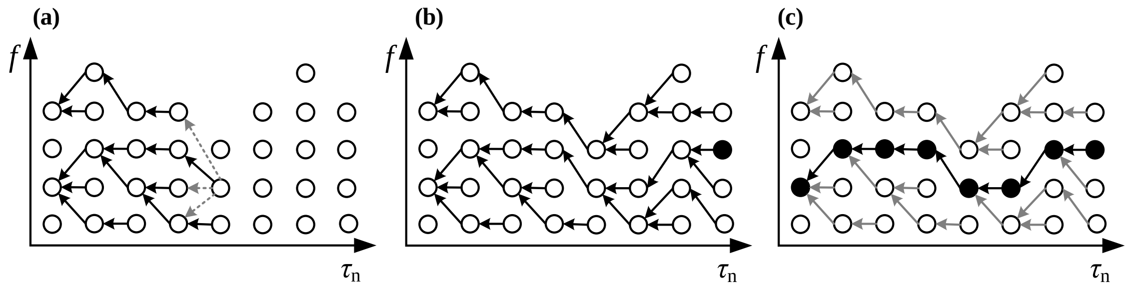

Fig. 5.1 Scheme of the T-F curve extraction (cited from [102]). (a) Step 1; (b) step 2; (c) step 3.

The solution of the fast path optimization is given as

$$\begin{aligned}
&\textit{for } n=1,\ldots,N \text{ , } m=1,\ldots,N_p(\tau_n) \text{ and } k=1,\ldots,N_p(\tau_{n-1}) \\
&q(m,\tau_n)=\arg\max_k\left\{F\left[Q_m(\tau_n),\nu_m(\tau_n),\nu_k(\tau_{n-1})\right]+U(k,\tau_{n-1})\right\},n>1 \\
&U(m,\tau_n)=\begin{cases} F\left[Q_m(\tau_n),\nu_m(\tau_n),\nu_m(\tau_n)\right] & n=1 \\ F\left[Q_m(\tau_n),\nu_m(\tau_n),\nu_{q(m,\tau_n)}(\tau_{n-1})\right]+U\left(q(m,\tau_n),\tau_{n-1}\right) & n>1 \end{cases}
\end{aligned} \tag{5.8}$$

where $q(m,\tau_n)$ indicates which peaks at the previous moment τ_{n-1} should be linked to the peaks at the current moment τ_n, and $U(m,\tau_n)$ is an intermediate variable which contributes to the optimization. The computational complexity is reduced to the order of $O\left(N_f N+M_p{}^2 N\log N\right)$, which is significantly lower than $O\left(N_f N+M_p{}^N\right)$. To further illustrate the T-F extraction approach, a scheme is shown in Fig. 5.1. The spots represent the peaks in the TFR determined via Eq.(5.2). The extraction is composed of three steps, shown as (a), (b) and (c) in Fig. 5.1, respectively. First, by applying the solution given in Eq.(5.8), $q(m,\tau_n)$ is calculated for all peaks from τ_2 to τ_N; then, the peak with the maximum value of $U(m,\tau_n)$ at τ_N is selected as the last point of the T-F curve, shown as the solid spot in Fig. 5.1(b); finally, from the selected point, the whole curve can be determined following the track. The extracted curve is noted as $f_p(\tau_n)$.

5.3.2 Multiple time-frequency curve extraction based on fast path optimization

Based on the fast path optimization approach for T-F curve extraction as introduced in the previous section, a multiple T-F curve extraction algorithm is proposed. The flowchart of the algorithm is shown in Fig. 5.2. Fast path optimization is iteratively applied to the TFR to extract multiple T-F curves. After one T-F curve is extracted, the peaks from the extracted T-F curve are removed from the current peak map to generate a new peak map. Then, with the new peak map, the next T-F curve can be extracted by applying fast path optimization again.

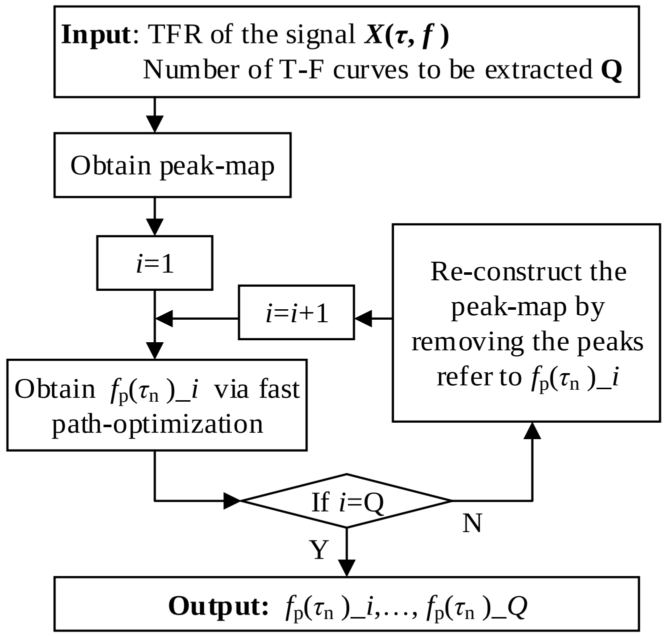

Fig. 5.2 Flowchart of multiple T-F curve extraction.

The number of curves to be extracted should be selected as sufficient to show the IFCF and at least one of its harmonics. Obviously, the more curves the better. However, more curves lead to higher computational cost. The computational complexity of the multiple T-F curve extraction algorithm is on the order of $O\left(QN_f N + QM_p{}^2 N \log N\right)$. Ideally, the Q value should be changed for different signals according to the composition of the signal and the intensity of the noise. Therefore, a guideline is that the Q should be high enough to reveal the IFCF and its harmonics, while the computational cost remains acceptable. In this chapter, Q=5 is selected for all the tests in simulations and experiments. However, different Q values can be selected for different cases, as long as the selection satisfies the guideline.

5.4 Proposed procedure for bearing fault diagnosis under unknown time-varying speed conditions

With the multiple T-F curve extraction algorithm, a procedure for bearing fault diagnosis under time-varying rotational speed conditions is proposed in the flowchart shown in Fig. 5.3. This procedure is resampling-free and tachometer-free.

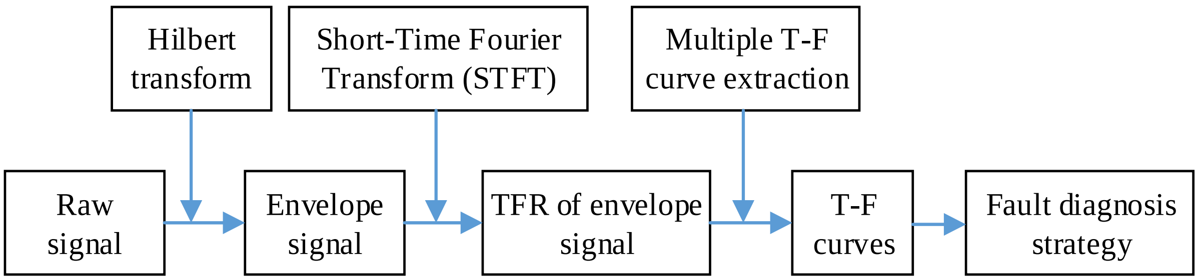

Fig. 5.3 Procedure for bearing fault diagnosis under unknown time-varying speed conditions.

The raw signal is first amplitude demodulated via the Hilbert transform, after which the demodulated signal, also called the envelope signal, is obtained. Then, the STFT is employed to obtain the TFR. Subsequently, the multiple T-F curve extraction algorithm is applied to the TFR to extract a certain number of T-F curves. A fault diagnosis strategy is proposed with the extracted T-F curves.

The flowchart of the fault diagnosis strategy is shown in Fig. 5.3. The Fault Characteristic Coefficient (FCC), defined as the constant ratio of the FCF to the Shaft Rotating Frequency (SRF) is utilized in the flowchart. The formulae for the FCF of outer race fault (BPFO) and inner race fault (BPFI) are given as [9]

$$\text{BPFO}=\frac{n_b}{2}\left(1-\frac{d}{D}\cos\phi\right)f_r=FCC_O\,f_r \tag{5.9}$$

$$\text{BPFI}=\frac{n_b}{2}\left(1+\frac{d}{D}\cos\phi\right)f_r=FCC_I\,f_r \tag{5.10}$$

where n_b is the number of rolling elements, d is the diameter of the rolling element, D is the pitch diameter of the bearing, ϕ is the angle of the load from the radial plane, f_r is the shaft rotational frequency, FCC_O and FCC_I are respectively the outer race and inner race FCCs. It is clear that the ratio of the FCF to the SRF is constant since it is determined by the dimensions of the bearing. Theoretically, for time-varying speed cases, the IFCF should maintain the same ratio with the ISRF. Based on this, if the extracted T-F curves cover the IFCF, its harmonics and the ISRF, then the fault can be detected and diagnosed by matching the average ratios of the curves to the extracted ISRF with FCC, i.e., the ratio of the FCF to f_r. The average ratio between two extracted curves $f_p(\tau_n)\_i$ and $f_p(\tau_n)\_j$ is defined in this chapter as

$$R_a = \frac{1}{N}\sum_{n=1}^{N}\frac{f_p(\tau_n)\_i}{f_p(\tau_n)\_j} \tag{5.11}$$

The bottom of the ratio is taken as the extracted T-F curve which has the lowest frequency since the ISRF is smaller than the IFCF and its harmonic. The ratio then becomes

$$R_a = \frac{1}{N}\sum_{n=1}^{N}\frac{f_p(\tau_n)\_i}{f_p(\tau_n)\_l} \tag{5.12}$$

where $f_p(\tau_n)\_l$ denotes the T-F curve with the lowest frequency. It should be pointed out that the calculated average ratio may not completely match the FCC due to the limited resolution of the TFR and the limited precision of the given dimensions of the bearing. Thus, it is proposed to use a 5% relative error of the calculated average ratio to the FCC for tolerance of the matching, i.e. $|R_a - \mathrm{FCC}| \bullet 100\% / \mathrm{FCC} \leq 5\%$. If the calculated average ratios of the T-F curves to the curve of the lowest frequency match the FCC and its harmonics (n*FCC), then the corresponding fault type is determined. Additionally, the sidebands with the frequency n*FCF± f_r may be also observed in the TFR. The average ratios of sidebands to f_r are n*FCC±1 since FCF=FCC* f_r.

However, it is possible that the ISRF may not be covered in the extracted Q curves. The ISRF shown in the TFR can be the frequency of the modulating signal of the bearing fault-induced signal [9] or the fundamental frequency of the signal caused by misalignment, eccentric or imbalanced mass [36]. If the bearing fault-induced signal is not amplitude modulated by the modulating signal with the ISRF as the fundamental frequency and the machine operates under a properly installed condition (e.g., with minimal misalignment, eccentric and imbalanced mass), the ISRF shown in the TFR can be relatively weak. Thus, an iterative process is included in the fault diagnosis strategy. The flowchart of the fault diagnosis strategy is shown in Fig. 5.4. If the average ratios of the extracted Q curves do not match the FCCs, which implies that the ISRF was not included in the Q extracted curves, then extra T-F curves need to be extracted. Each new curve is extracted from the portion below the previously extracted curve of the lowest frequency in the TFR. With the newly extracted T-F curve, the average ratios can be recalculated. If the ratios are still not matched with the FCCs, then the new T-F curve is extracted iteratively until the repeat times reaches a chosen limit P_{max}. Similar to the setting of Q, P_{max} has to be sufficient enough to ensure that the ISRF is extracted if the bearing is faulty. However, it cannot be too large because of the increased computational cost. In this chapter, P_{max}=3 was selected in the simulations and experiments. The bearing is considered healthy if no ratio matches one of the FCCs after the complete process.

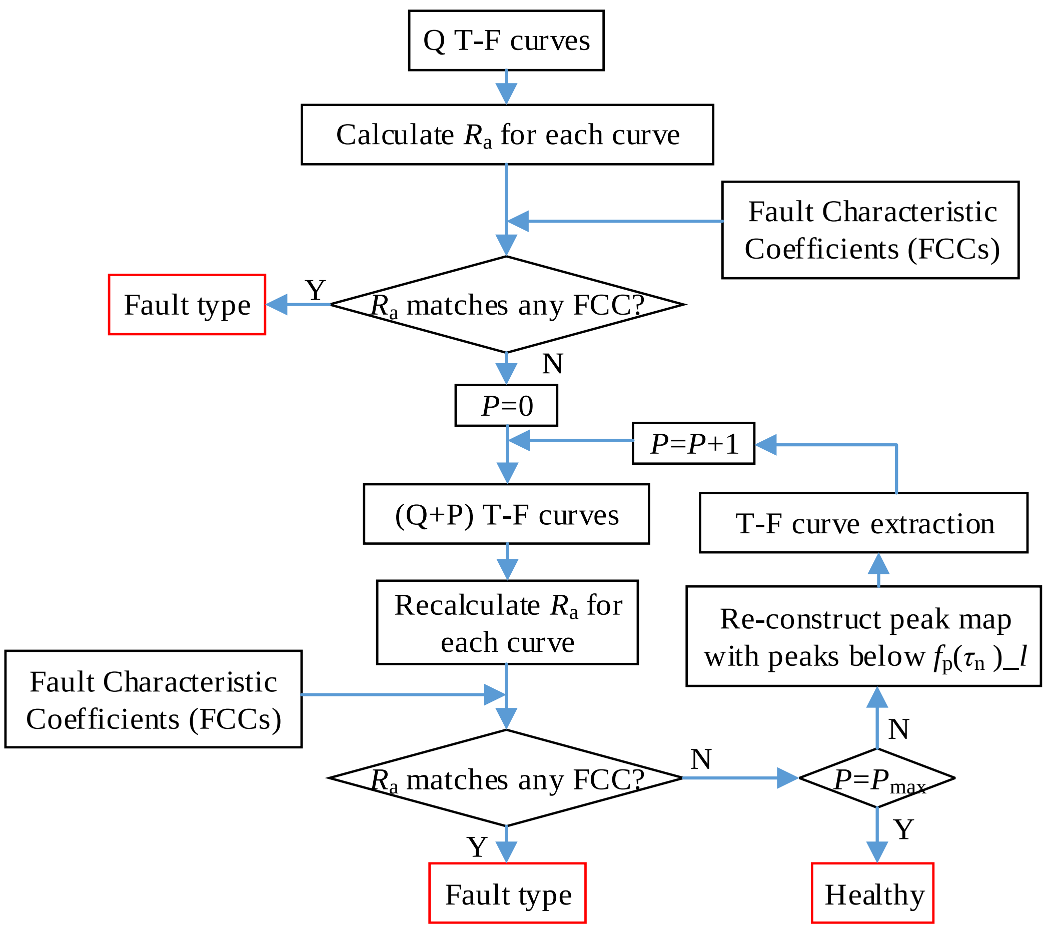

Fig. 5.4 Fault diagnosis strategy.

5.5 Simulations

Simulations were performed in order to illustrate the application of the proposed procedure for bearing fault diagnosis under unknown time-varying speed conditions and also to test the performance of the procedure.

The vibration signal induced by a bearing defect can be simulated as impulse responses of a 1 degree-of-freedom mass-spring-damper system, expressed as [36]

$$x(t)=\sum_{m=1}^{M} L_m \mathrm{e}^{-\beta\left(t-mT_p-\sum_{i=1}^{m}\varepsilon_i\right)} \sin\left[\omega_r\left(t-mT_p-\sum_{i=1}^{m}\varepsilon_i\right)\right] u\left(t-mT_p-\sum_{i=1}^{m}\varepsilon_i\right) \qquad (5.13)$$

where M is the number of the impulse responses, L_m is the amplitude coefficient, T_p is the period of the impulses, the reciprocal of the FCF, β represents the damping characteristic determined by the structure of the system, ω_r is the damped natural frequency of the system, ε_i is the slipping coefficient, which is randomly simulated as $0.01T_p$-$0.02T_p$, and $u(t)$ is the unit step function. This model can be considered as an amplitude modulated signal with ω_r being the frequency of the carrier frequency and the FCF being the frequency of the modulating signal. The demodulated signal is called the envelope. One of the most commonly used demodulation methods is the Hilbert transform. For a signal $x(t)$, the Hilbert transform is defined as [104]

$$H\left[x(t)\right]=\frac{1}{\pi}\int_{-\infty}^{\infty} x(\tau)\frac{1}{t-\tau}\mathrm{d}\tau \qquad (5.14)$$

With the envelope, it shows peaks at the FCF and its harmonics in the envelope spectrum. However, for the time-varying speed cases, T_p is also varying. In addition, the signal is always masked by noise. Also, the signal is considered further amplitude modulated with the ISRF as the frequency of the modulating signal if the bearing has an inner race fault [9]. Moreover, the collected vibration signal usually comprises the signal with the shaft rotational frequency as the fundamental frequency caused by misalignment, eccentric mass or imbalance [93]. To account for these four factors, the model can be modified to

$$x(t) = A(t)\sum_{m=1}^{M} L_m \mathrm{e}^{-\beta(t-t_m)} \sin\left[\omega_r\left(t-t_m\right)\right] u\left(t-t_m\right) + s\left(t\right) + n(t) \tag{5.15}$$

where $A(t) = 1 + \alpha\cos\left(2\pi f_r\left(t\right)t\right)$ is the factor used to reflect the amplitude modulation of the responses with the ISRF ($\alpha > 0$ for inner race fault and $\alpha = 0$ for outer race fault), $f_r\left(t\right)$ being the frequency of the modulating signal; α is a constant; $L_m = L_0 + \eta f_r\left(t_m\right)$ is the amplitude which is proportional to the rotational speed; L_0 and η are constants, t_m is the occurrence time of the *m*th impulse, $s\left(t\right)$ is the sinusoidal signal with the shaft rotational frequency as the fundamental frequency, and $n\left(t\right)$ is random noise, SNR can be used to describe the intensity of the noise. The occurrence time of the impulses can be computed by

$$\begin{cases} t_1 = \left(1+\mu\right)\left[1/f_c\left(0\right)\right] \\ t_m = \left(1+\mu\right)\left[1/f_c\left(0\right) + 1/f_c\left(t_1\right) + \ldots + 1/f_c\left(t_{m-1}\right)\right] \quad m = 2,3,\ldots,M \end{cases} \tag{5.16}$$

where μ is the slippage ratio varying from 0.01 to 0.02, $f_c\left(t\right)$ denotes the IFCF, the time interval between two impulses is $1/f_c\left(t\right)$ and $f_c\left(0\right)$ cannot be zero. $s\left(t\right)$ can be simulated by

$$s\left(t\right) = \sum_{i=1}^{N_s} B_i \cos\left(2\pi f_r\left(t\right)\cdot t\right) \tag{5.17}$$

where N_s is the number of harmonics of $s\left(t\right)$ and B_i is the amplitude of i^{th} harmonic of $s\left(t\right)$. With the demodulated signal, the IFCF and its harmonics show curves in the TFR of the envelope signal if $\alpha \neq 0$. The rotational frequency and sidebands $f_c\left(t\right) \pm n f_r\left(t\right)$, $n = 0,1,\ldots$ also show curves in the TFR.

To test the performance of the proposed method for bearing fault diagnosis under time-varying speed conditions, different types of signals are simulated. First, both linear and non-linear speed variations are considered. For linear variation cases, speed-up is utilized for the simulation since speed-up and speed-down follow the same pattern. For the non-linear cases, exponential decay speed and quadratic speed with speed-up plus speed down are considered. Additionally, to test the fault diagnosis strategy, two conditions featuring respectively dominant

ISRF and non-dominant ISRF in the TFR are simulated for the linear speed variation cases with the different levels of clearness of the ISRF in the TFR.

5.5.1 Simulated signal with linear speed-up ISRF

5.5.1.1 Case 1: ISRF is dominant in the TFR

A signal with linearly speed-up ISRF is simulated to test the performance of the proposed method, where $f_r(t)$=20+3t. The parameters in Eq.(5.15) are set as α =0.9, L_0 =1, η =0.05, β =500, ω_r =2π*4000 rad/s, μ =0.01, N_s =2, B =[0.8, 0.2], and SNR = -5dB. Additionally, the FCC is set as 3.5, i.e. $f_c(t) = 3.5 f_r(t)$. The sampling frequency of the signal is 20 kHz and the signal length is 10s.

The simulated signal is shown in Fig. 5.5(a). It can be seen that the amplitude of the signal increases with the increase of the ISRF. Also, the spacing of the impulses becomes smaller due to the increasing ISRF. By applying the Hilbert transform and the STFR to the simulated signal, the TFR of the envelope signal is obtained as shown in Fig. 5.5(b). From the TFR, many ridges can be observed, including the ISRF, IFCF, its harmonics, as well as the sidebands. The ridge of the ISRF is almost as clear as the ridge of the IFCF.

According to the proposed procedure for fault diagnosis, the multiple T-F curve extraction algorithm is applied to the TFR first. The number of curves to be extracted, Q, is set as 5 in this simulation, thus five curves are extracted shown in Fig. 5.5(c). The corresponding average ratio Ra of each curve is calculated as 1, 2.50, 3.50, 4.50 and 6.99, respectively. Among the average ratios, 3.50 matches the FCC (3.5) completely and 6.99 matches 2*FCC. Additionally, 2.50 and 4.50 match FCC-1 (sideband) and FCC+1 (sideband), respectively. Therefore, the fault diagnosis is completed and the extracted curve of the lowest frequency is the ISRF and the one of the third lowest is the IFCF.

Comparisons of the extracted ISRF and IFCF to their pre-set curves are shown in Fig. 5.5(d). It can be seen that both the extracted ISRF and IFCF are well matched with the pre-set ones. The average relative error of the extracted IFCF to the calculated IFCF is 0.51% and the average relative error of the extracted ISRF to the preset ISRF is 1.52%, respectively.

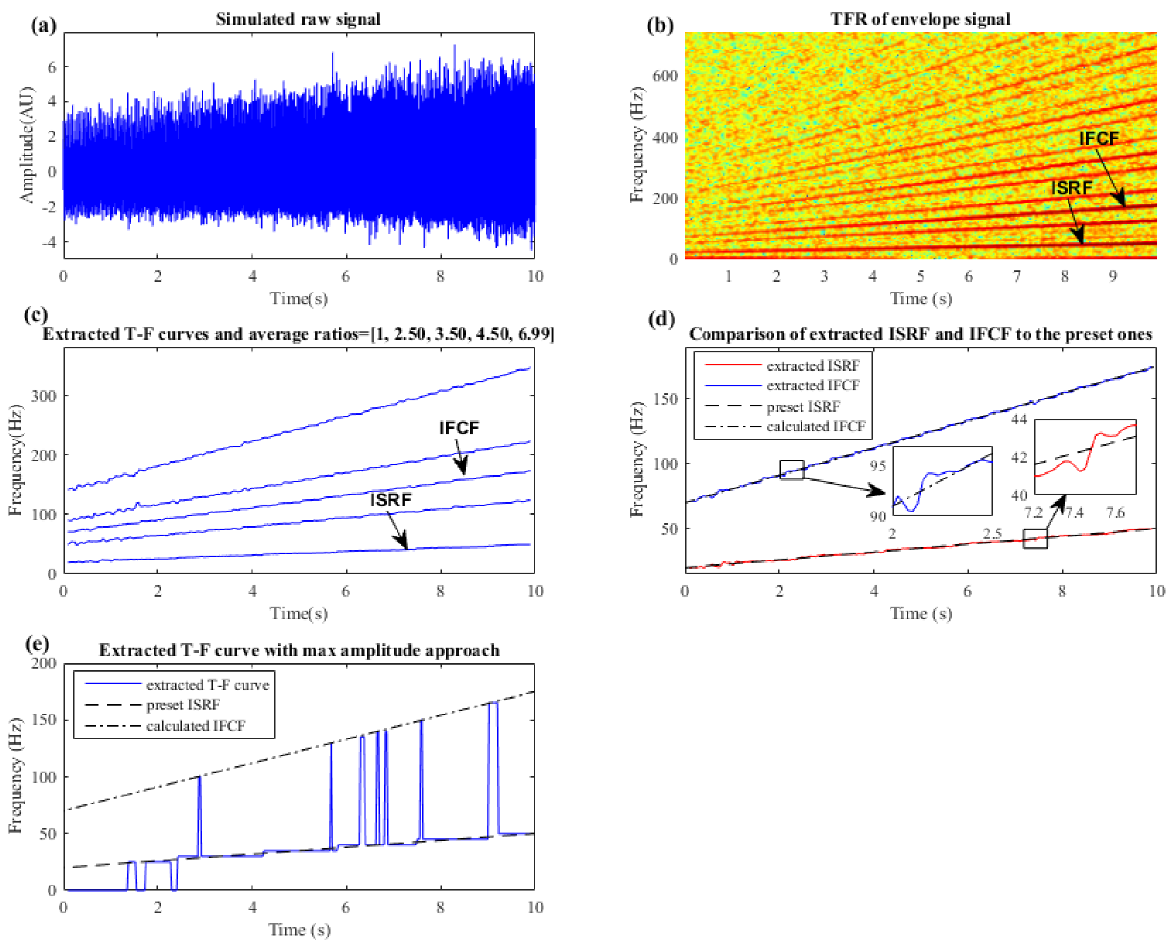

Fig. 5.5 Results of the proposed method applied to the simulated signal with linear speed-up ISRF and dominant ISRF in the TFR. (a) Simulated raw signal; (b) TFR of envelope signal; (c) extracted T-F curves; (d) comparison of extracted ISRF and IFCF to their preset curves; (e) extracted T-F curve by max amplitude approach.

To compare the fast path-optimization approach with the maximum amplitude approach for T-F curve extraction, the extracted T-F curve by extracting the ridge point of the maximum amplitude at each moment is shown in Fig. 5.5(e). The preset ISRF and calculated IFCF are also shown in the same figure. It can be seen that the extracted T-F curve is quite messy, jumping from the bottom to the IFCF, and then back and forth between the IFCF and ISRF. Obviously, the result is not as reliable as the curves shown in Fig. 5.5(c) obtained via the proposed multiple T-F curve extraction algorithm.

The performance of the proposed method to the simulated signal blended with different levels of noise is also tested. Due to space limit, only the average relative errors of the IFCF and the average relative errors of the ISRF are reported in this chapter as shown in Table 5.1, where SNR varies from SNR=0dB to SNR=-10dB. For each SNR, the average relative error of the IFCF is smaller than the average relative error of the ISRF. The reason is that the IFCF of the simulated signal is more significant than the ISRF in the TFR. In addition, with the increase of the noise level, the average relative errors of the IFCF and the ISRF both increase. The average relative error of the IFCF is increased by 0.41% from 0.44% to 0.85% and the average relative error of ISRF is increased by 2.5% from 1.41% to 3.91% when SNR is reduced from 0dB to -10dB. The results when SNR = -10dB are still acceptable. Also, the increase rates of the average relative errors for both the IFCF and the ISRF become larger with the decrease of SNR.

Table 5.1 Average relative errors of IFCF and ISRF with different level of noise for case 1.

	SNR=0dB	SNR=-3dB	SNR=-5dB	SNR=-8dB	SNR=-10dB
Average relative error (IFCF)	0.44%	0.49%	0.51%	0.62%	0.85%
Average relative error (ISRF)	1.41%	1.48%	1.52%	2.39%	3.91%

5.5.1.2 Case 2: ISRF is weak in the TFR

To test the performance of the proposed procedure in the case that the ISRF is not extracted from the TFR by the multiple T-F curve extraction algorithm, a signal with a weak ISRF is simulated in this section. The rotational frequency is kept the same at $f_r(t)$=20+3*t*, and the parameters are set as α=0, L_0=1, η=0.05, β=500, ω_r=2π*4000 rad/s, μ=0.01, N_s=2, B=[0.7, 0.2], SNR= -5dB, and FCC=3.5, i.e. $f_c(t)=3.5f_r(t)$. Compared to the settings for case 1, α is smaller, which leads to less visible ridges of the ISRF in the TFR. The sampling frequency of the signal is 20 kHz and the signal length is 10s.

The simulated signal is shown in Fig. 5.6(a). The amplitude of the signal is increasing with the increase of the rotational frequency. By applying the STFR to the demodulated signal, the TFR is obtained as shown in Fig. 5.6(b). The IFCF and its harmonics can be observed from the TFR. Unlike the TFR shown in Fig. 5.5(b), the ISRF is not very clearly shown in the TFR, especially in the low-frequency portion.

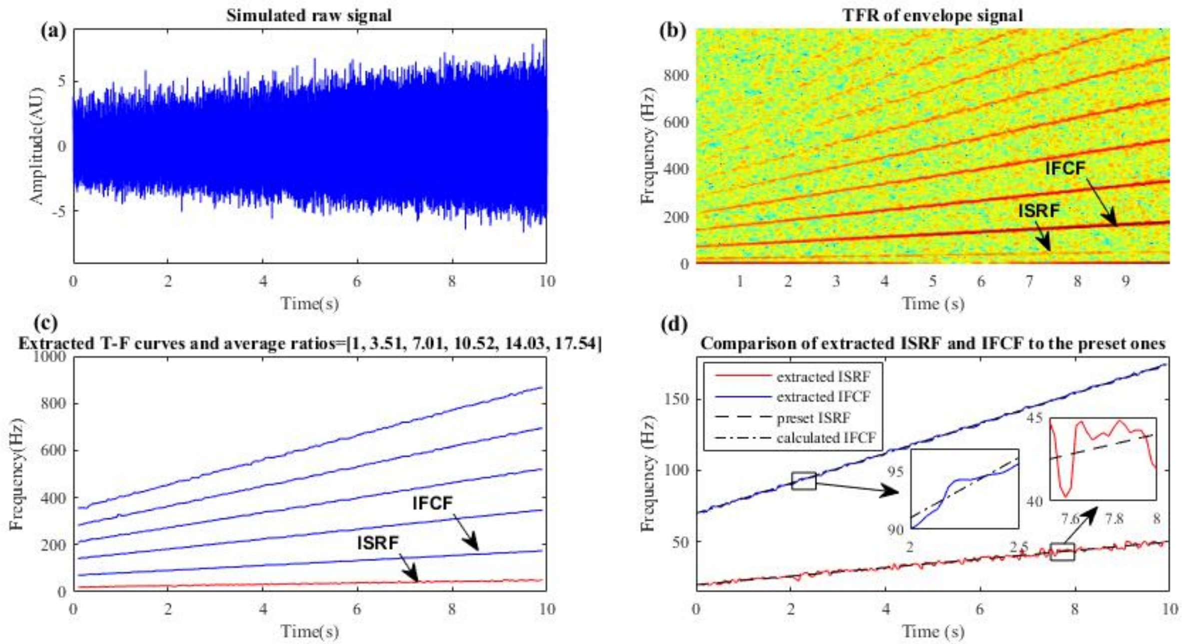

Fig. 5.6 Results of the proposed method applied to the simulated signal with linear speed-up ISRF and weak ISRF in the TFR. (a) Simulated raw signal; (b) TFR of envelope signal; (c) extracted T-F curves; (d) comparison of extracted ISRF and IFCF to their preset curves.

The extracted T-F curves are shown in Fig. 5.6(c) by applying the proposed procedure to the simulated signal. First, the multiple T-F curve extraction algorithm is used to extract curves. The Q is again set as 5, thus five curves are extracted, shown in blue in Fig. 5.6(c). The average ratios of the curves to the one of lowest frequency are calculated as 1, 2, 3, 4, and 5, respectively. None of the ratios matches the FCC (3.5). Therefore, according to the fault diagnosis strategy, an additional curve is extracted from the TFR below the ridge of the lowest frequency previously extracted. Here, P_{max} is set as 3 for the fault diagnosis strategy. The newly extracted curve is shown in red in Fig. 5.6(c). With the new curve, the average ratios are recalculated as 1, 3.51, 7.01, 10.52, 14.03 and 17.54. Among the recalculated average ratios, 3.51 matches the FCC (3.5) within 5% relative error. Also, average ratios 7.01, 10.52, 14.03 and 17.54 are matched with 2*FCC, 3*FCC, 4*FCC and 5*FCC, respectively, which implies the 2nd, the 3rd, the 4th and the 5th harmonics of the IFCF. Therefore, the fault diagnosis can be completed.

A comparison diagram of the extracted IFCF and the extracted ISRF to the preset curves is shown in Fig. 5.6(d). The average errors of the IFCF and the ISRF are calculated as 0.46% and 2.78%, respectively. It can be seen that the extracted IFCF matches the calculated IFCF. The extracted ISRF fits the pre-set ISRF overall, however, does not match well in the low-frequency portion. The reason is that the ridge of the ISRF in the relatively low-frequency portion is less clear in the TFR.

The proposed method is also applied to the signal with different levels of noise. The average relative errors of the IFCF and the average relative errors of the ISRF are given in Table 5.2. Similar to case 1, for each SNR, the average relative errors of the ISRF are higher than those of the IFCF. The reason is that the ridge of the IFCF is significantly clearer than that of the ISRF in the TFR. The average relative errors of both the IFCF and the ISRF increase with the increase in noise level. The average relative error of the IFCF increases by 0.26% from 0.39% to 0.65% and the average relative error of ISRF increases by 8.41% from 1.71% to 9.85% when the SNR is reduced from 0dB to -10dB. The results when SNR=-10dB are still acceptable. Additionally, compared to the results given in Table 5.1, the average relative errors of the IFCF for case 2 is smaller than those for case 1. However, the average relative errors of the ISRF for case 2 are higher than those for case 1. The reason can be found from the comparison of the clearness of the ISRF and the IFCF in Fig. 5.5(b) to the clearness of the ISRF and the IFCF in Fig. 5.6(b), respectively. It can be seen that the ridge of the IFCF in Fig. 5.6(b) is clearer than the one in Fig. 5.5(b), however, the ridge of the ISRF in Fig. 5.6(b) is less clear than the one in Fig. 5.5(b).

Table 5.2 Average relative errors of ISRF and IFCF with different level of noise for case 2.

	SNR=0dB	SNR=-3dB	SNR=-5dB	SNR=-8dB	SNR=-10dB
Average relative error of IFCF	0.39%	0.44%	0.46%	0.54%	0.65%
Average relative error of ISRF	1.71%	1.99%	2.78%	6.98%	9.85%

5.5.2 Simulated signal with non-linear ISRF

5.5.2.1 Case 3: Signal with exponentially decaying ISRF

To test the performance of the proposed method using a bearing signal with non-linearly varying rotational speed, the signal with an exponentially decaying ISRF is simulated in this section. The rotational frequency is set as $f_r(t)=80e^{-0.35t}$ and the other parameters are α =0.9, L_0 =1, η =0.05, β =500, ω_r =2π*4000 rad/s, μ =0.01, N_s =2, B =[0.8, 0.2], SNR= -5dB, and FCC=3.5, i.e. $f_c(t)=3.5f_r(t)$. The sampling frequency of the signal is 20 kHz and the signal length is 5s.

The simulated signal with the given settings is shown in Fig. 5.7(a). It can be seen that with the decreasing speed, the density of the signal reduces and the amplitude of the signal also reduces. The TFR of the envelope of the simulated signal is shown in Fig. 5.7(b), obtained via the STFT. Similar to case 1, the IFCF and its harmonics, the ISRF and the sidebands show ridges in the TFR. Additionally, the ISRF is almost as clear as the IFCF.

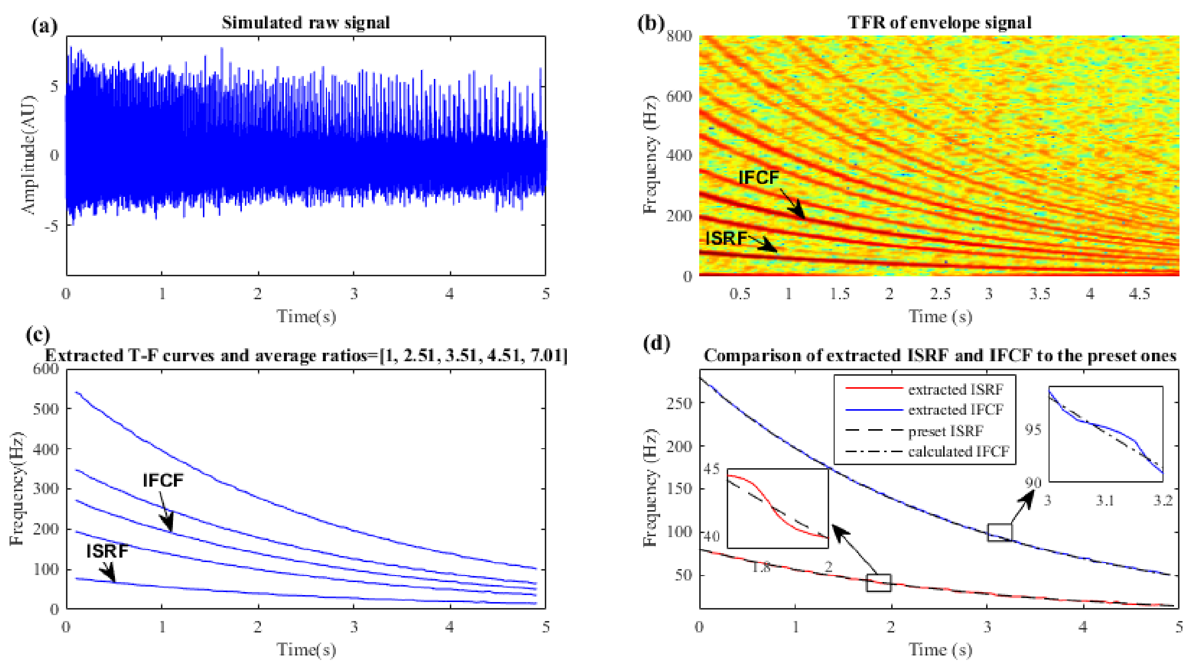

Fig. 5.7 Results of the proposed method applied to the simulated signal with exponentially decaying ISRF. (a) Simulated raw signal; (b) TFR of envelope signal; (c) extracted T-F curves; (d) comparison of extracted ISRF and IFCF to their preset curves.

Applying the proposed multiple T-F curve extraction algorithm to the TFR with Q=5, five T-F curves are extracted, as shown in Fig. 5.7(c). The average ratios are calculated as 1, 2.51, 3.51, 4.51 and 7.01, respectively. Obviously, among those average ratios, 3.51 matches the FCC of (3.5) within a 5% relative error and 7.01 matches 2*FCC. In addition, 2.51 and 4.51 match the coefficient for the sidebands (FCC-1 and FCC+1). Therefore, according to the proposed procedure, the fault diagnosis is completed.

Compared to the curve at which the average ratio is 1 to the pre-set ISRF and the curve at which the average ratio is 3.51 to the calculated IFCF, they are closely overlapped, seen in Fig. 5.7(d). The average relative errors of the IFCF and the ISRF are 0.46% and 1.69%, respectively.

The proposed procedure is also applied to the signal with different levels of noise. The average relative errors of the IFCF and the average relative errors of the ISRF are given in Table 5.3, where the SNR varies from 0dB to -10dB. The average relative errors of the ISRF are higher than those of the IFCF for each SNR except SNR = -10dB. Additionally, the average relative errors of both the ISRF and the IFCF increase with the decrease of the SNR. When SNR decreases from 0dB to -8dB, the average relative error of the IFCF increases by 0.12% from 0.41% to 0.53% and the average relative error of the ISRF increases slightly (0.26%) from 1.49% to 1.75%. However, when the noise further increases to SNR = -10dB, average relative errors of the IFCF and the ISRF become 15.38% and 7.82%, which is beyond the acceptable limits. This implies that the proposed method for an exponentially decaying ISRF case has less tolerance to noise compared to the cases with linearly varying ISRF.

Table 5.3 Average relative errors of IFCF and ISRF with different level of noise for case 3.

	SNR=0dB	SNR=-3dB	SNR=-5dB	SNR=-8dB	SNR=-10dB
Average relative error of IFCF	0.41%	0.44%	0.46%	0.53%	15.38%
Average relative error of ISRF	1.49%	1.53%	1.69%	1.75%	7.82%

5.5.2.2 Case 4: Signal with quadratic ISRF

To further test the performance of the proposed method in processing the bearing signal with non-linearly varying rotational speed, a signal with quadratic ISRF is simulated in this section. The rotational frequency is set as $f_r(t)$=9.6t^2+48t+20 and the other parameters are α =0.9, L_0=1, η=0.1, β=500, ω_r=2π*4000 rad/s, μ=0.01, N_s=2, B=[0.8, 0.2], SNR = -5dB, and FCC=3.5, i.e. $f_c(t)=3.5f_r(t)$. The sampling frequency of the signal is 20 kHz and the signal length is 5s.

The simulated signal with the given settings is shown in Fig. 5.8(a). With quadratically varying rotational speed, the amplitude of the signal increases in the left half time span and then reduces in the right half time span. The TFR of the envelope of the simulated signal is shown in Fig. 5.8(b), obtained via the STFT. It can be seen that the IFCF and its harmonics, ISRF and sidebands show ridges in the TFR. Similarly, the ISRF is almost as clear as the IFCF. However, for high order harmonics, the ridges at the low-frequency portion are unclear.

According to the proposed procedure, multiple T-F curve extraction is applied to the TFR with results shown in Fig. 5.8(c). It can be observed that extracted curves of the highest frequency and the second highest frequency exhibit some meandering pattern at the low-frequency portion, i.e. at the beginning and at the end. This is due to the unclear ridges at the corresponding portion in the TFR. The average ratios of the five extracted T-F curves are calculated as 1, 2.50, 3.49, 4.49, and 6.99, respectively. Among these average ratios, 3.49 matches the FCC (3.5) within 5% relative error and 6.99 matches 2*FCC. In addition, 2.50 and 4.49 match FCC-1 and FCC+1 (i.e. sidebands), respectively. Then, the extracted curve of the lowest frequency is considered as the ISRF and the curve with average ratio 3.49 is identified as the IFCF. With these results, the fault can be diagnosed even though two of the extracted curves are less ideal.

Comparisons of the extracted IFCF and the extracted ISRF to the preset curves are shown in Fig. 5.8(d). Also, two local comparisons (in the time duration of 0 to 1 second) between the extracted IFCF and the calculated IFCF, and between the extracted ISRF and the preset ISRF are provided in Fig. 5.8(e) and Fig. 5.8(f), respectively, to reveal the close-up views. It can be seen that the extracted ISRF and the extracted IFCF are almost identical to the pre-set curves. The average relative error of the IFCF is 0.42% and the average relative error of the ISRF is 1.34%, respectively.

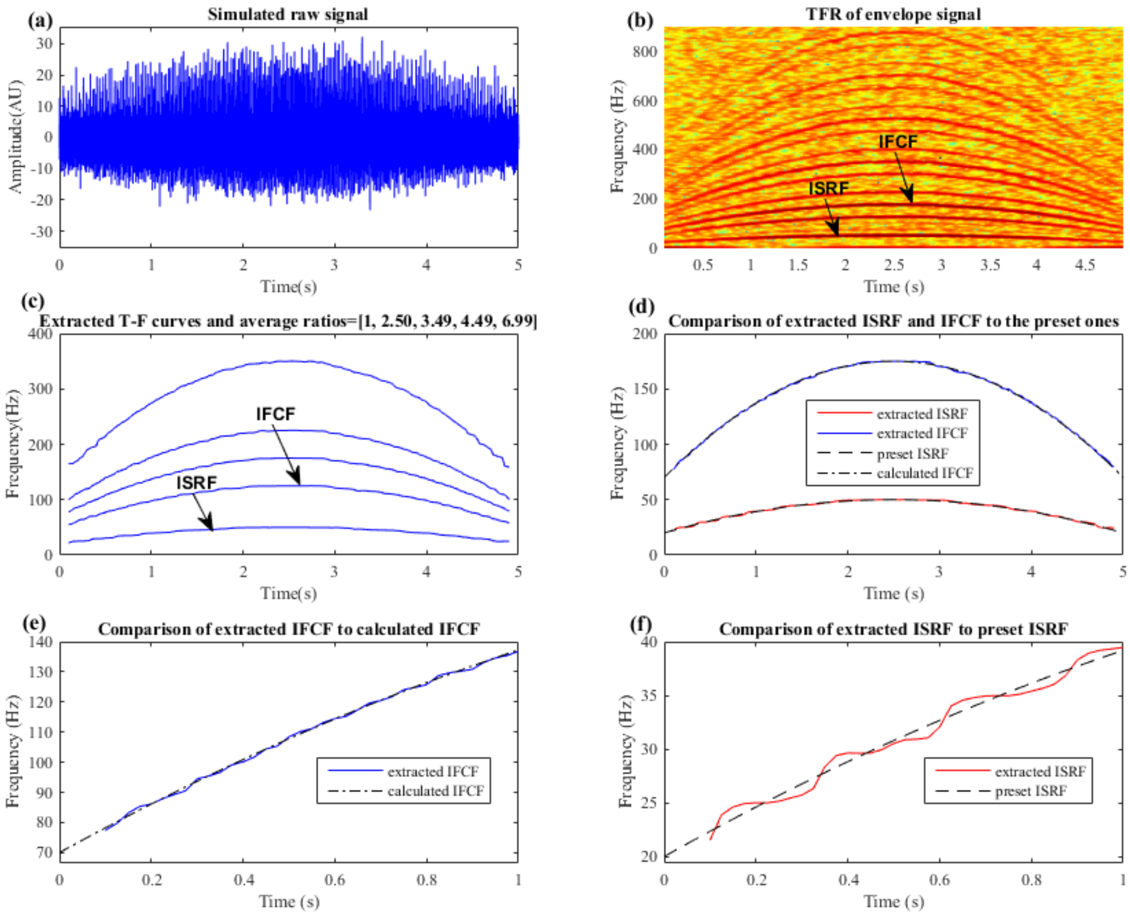

Fig. 5.8 Results of the proposed method applied to the simulated signal with quadratic ISRF. (a) Simulated raw signal; (b) TFR of envelope signal; (c) extracted T-F curves; (d) comparison of extracted ISRF and IFCF to their preset curves; (e) comparison of extracted IFCF to calculated IFCF (local view); (f) comparison of extracted ISRF to calculated ISRF (local view).

The performance of the proposed method is also tested on the signals with different SNR for the quadratic ISRF. The results of the average relative errors of the IFCF and the ISRF are given in Table 5.4 when SNR varies from 0dB to -10dB. Similar to case 3, the average relative errors of the ISRF are higher than those of the IFCF for each SNR except SNR = -10dB. Additionally, the average relative errors of the IFCF and the ISRF both increase with a decrease of the SNR. When the SNR reduces from 0dB to -8dB, the average relative error of the IFCF increases 1.71% from 0.35% to 1.06% and the average relative error of the ISRF increases 1.01% from 1.06% to 2.07%. The increase rate is higher than for case 3 but the results are still acceptable. However, when the noise further increases to SNR=-10dB, the results become unacceptable with average relative errors of the IFCF 12.35% and the ISRF 8.24%, which implies a narrower tolerance range of the proposed method when applied to the signal with quadratic ISRF.

Table 5.4 Average relative errors of IFCF and ISRF with different level of noise for case 4.

	SNR=0dB	SNR=-3dB	SNR=-5dB	SNR=-8dB	SNR=-10dB
Average relative error of IFCF	0.35%	0.40%	0.42%	1.06%	12.35%
Average relative error of ISRF	1.06%	1.16%	1.34%	2.07%	8.24%

The results of simulation illustrate that the proposed procedure for bearing fault diagnosis under unknown time-varying rotational speed conditions is effective regardless whether the varying rotational speed is either dominant or non-dominant in the TFR, and it is also applicable to both linear and non-linear speed variation patterns. However, its performance is less robust to noise in the non-linearly varying rotational speed cases in comparison to the linearly varying rotational speed cases. Furthermore, the results of the T-F curve extraction are also affected by the quality, i.e., clearness of the ridges in the TFR.

5.5.3 Data processing time

To provide more information about the computational cost of this method, the processing times required in the simulations are given in Table 5.5. All the simulations were performed using MATLAB R2015b, on a computer running Windows 7 with an Intel Core i5-4690 3.5 GHz processor and 8.0 GB of RAM. The results in Table 5 show that the computational cost in processing simulation data is affected by the signal sampling rate, signal length, number of extracted curves, and the time-varying pattern of the ISRF.

Table 5.5 Processing time required for simulation data analysis.

	Sampling rate	Signal length	Number of extracted curves	Processing time	Average processing time for 1 sec of data
Case 1	20kHz	10s	5	21.08s	2.11s
Case 2	20kHz	10s	6	22.60s	2.26s
Case 3	20kHz	5s	5	11.90s	2.38s
Case 4	20kHz	5s	5	10.00s	2.00s

5.6 Experimental validation

To further validate the effectiveness of the proposed method for bearing fault diagnosis under unknown time-varying speed conditions, experiments are conducted to collect vibration signals from bearings operating under varying speed. The experimental set-up is shown in Fig. 5.9. The shaft is driven by a motor which is running at variable speed, controlled by the AC driver. Two bearings are mounted to support the shaft, the one on the right side is healthy and the one on the left side is changed with three different bearings for separate experiments, i.e. a bearing with an outer race defect, a bearing with an inner race defect and a healthy bearing. An accelerometer is installed on the housing of the bearing on the left side to collect vibration signals. To validate the result of the proposed method, a tachometer is used to measure the shaft rotating speed. The measured signals are collected by a DAQ card and sampled using LABVIEW.

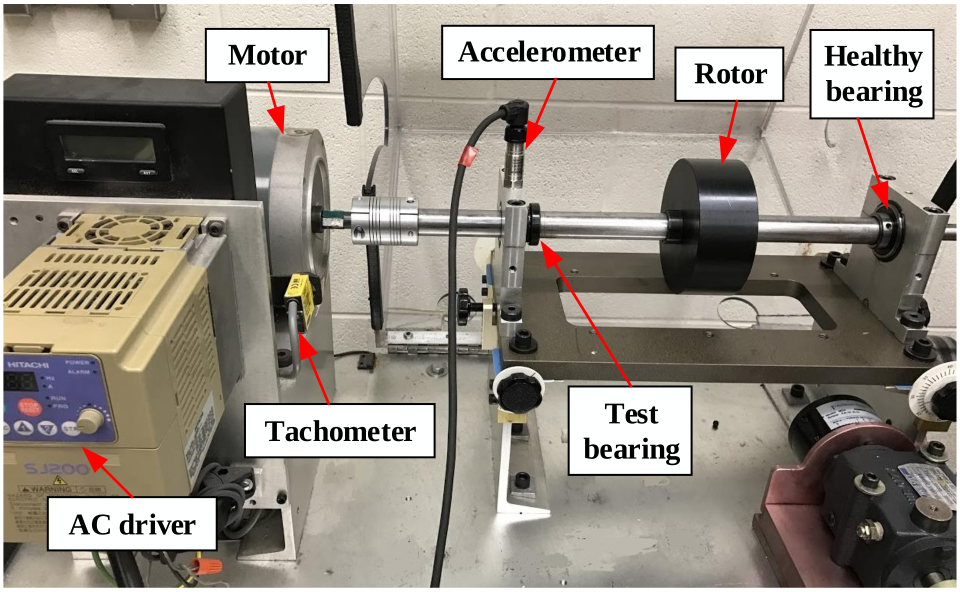

Fig. 5.9 Experimental set-up.

5.6.1 Bearing with outer race fault

The experiment is first conducted on a bearing with an outer race fault, i.e. the left bearing shown in Fig. 5.9 is a bearing with an outer race defect. The parameters of the bearing used for the experiment are given in Table 5.6. According to Eq.(5.9), the FCC of the bearing for outer race fault is calculated as 3.05. The sampling frequency of the signal is set as 24kHz, and the length of the signal is 5.3s.

Table 5.6 Parameters of bearings used for outer race fault test.

Bearing type	Pitch diameter	Ball diameter	Number of balls	FCF
ER10K	33.50mm	7.94mm	8	$3.05f_r$

The collected vibration signal and shaft rotational speed (i.e. ISRF) are shown in Fig. 5.10(a) and Fig. 5.10(b), respectively. The rotational speed is approximately linearly increasing from 25Hz to 40Hz. The IFCF can be calculated according to IFCF=FCC*ISRF. It can be seen from Fig. 5.10(a) that the spacing of the impulses becomes smaller and the amplitude of the signal becomes larger with the increase of the shaft rotational speed. The TFR of the amplitude demodulated signal is shown in Fig. 5.10(c), obtained via the STFT. Many ridges are shown in the TFR, including the ISRF. However, the IFCF cannot be clearly observed.

Applying the multiple T-F curve extraction algorithm to the TFR, the five curves are obtained, shown in Fig. 5.10(d). According to the proposed fault diagnosis procedure, the average ratios of the curves to the curve of the lowest frequency are calculated as 1, 3.01, 4.99, 6.03, and 9.04, respectively. Among these average ratios, 3.01 matches the FCC (3.05) with 1.3% relative error (within 5%). Additionally, 6.03 and 9.04 match 2*FCC and 3*FCC with relative error 1.1% and 1.2%, respectively. Also, 4.99 matches 2*FCC-1 (sideband) with relative error 2.2%. Therefore, an outer race fault can be diagnosed according to the proposed bearing fault diagnosis procedure. In addition, the extracted curve of the lowest frequency is revealed as the ISRF and the curve of the average ratio 3.01 is considered as the IFCF.

Comparisons of the extracted IFCF to the calculated IFCF and the extracted ISRF to the measured ISRF are shown in Fig. 5.10(e). It can be seen that the extracted ISRF and the extracted IFCF are in agreement with the measured ISRF and the calculated IFCF, respectively. The average relative errors of the IFCF and the ISRF are 2.11% and 5.45%, respectively.

To compare the fast path-optimization approach with the maximum amplitude approach for T-F curve extraction, the extracted T-F curve by extracting the point of the maximum amplitude at each moment is shown in Fig. 5.10(f). It can be seen that the curve is sharply discontinuous, which matches neither the IFCF, nor the ISRF. Obviously, the quality of extracted T-F curves shown in Fig. 5.10(d) is much better than that in Fig. 5.10(f).

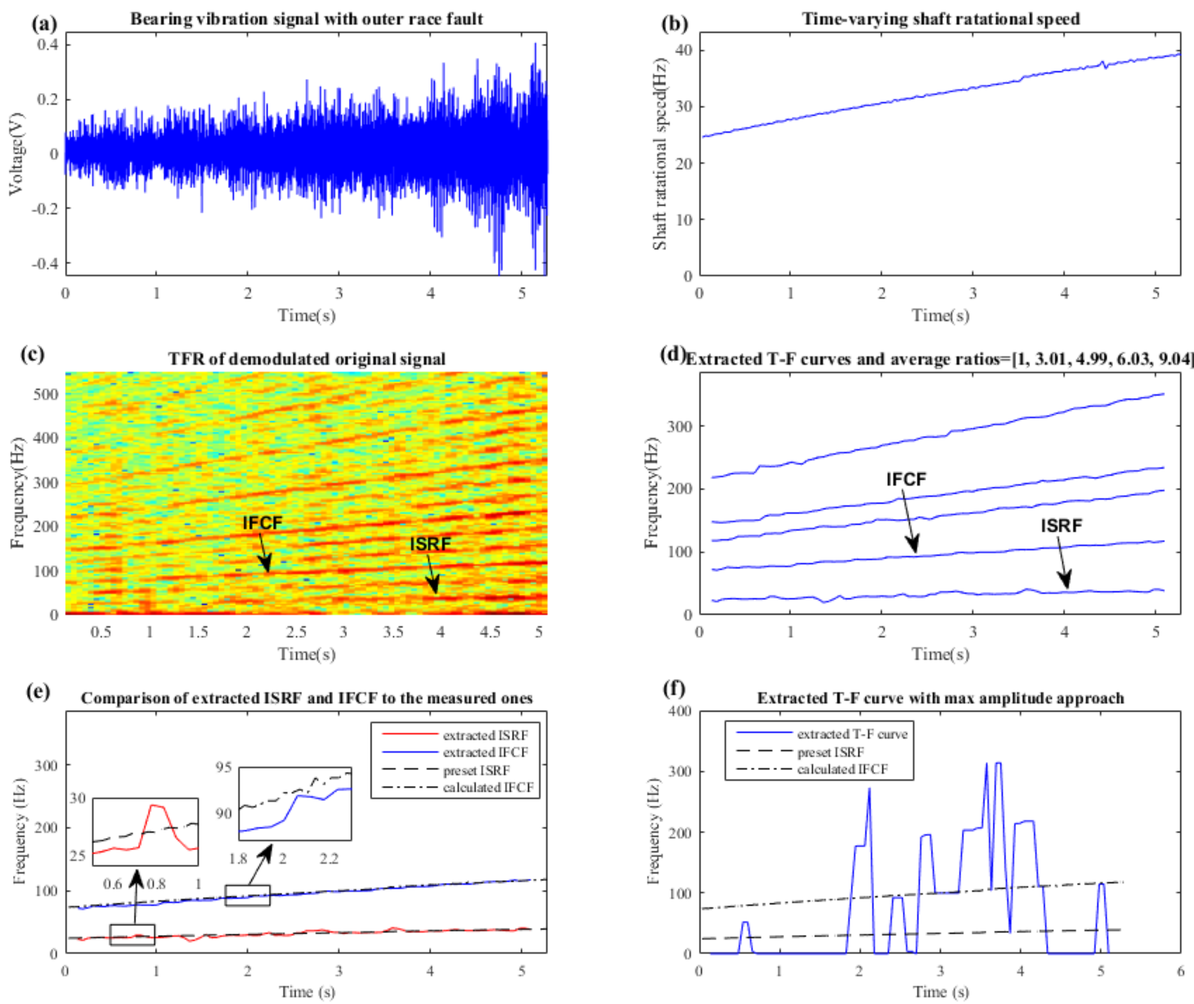

Fig. 5.10 Results of bearing with outer race fault. (a) Bearing vibration signal with outer race fault; (b) time-varying shaft rotational speed; (c) TFR of demodulated original signal; (d) extracted T-F curves; (e) comparison of extracted ISRF and IFCF to the measures curves; (f) extracted T-F curve with max amplitude approach.

5.6.2 Bearing with inner race fault

The experiment is also conducted by replacing the left bearing shown in Fig. 5.9 with one having an inner race defect. The parameters of the bearings used in the test are given in Table 5.7. According to Eq.(5.10), the FCC of the inner fault is calculated as 5.43 with the given parameters. The sampling frequency of the signal is set as 12 kHz and the length of the signal is 17.9s.

Table 5.7 Parameters of bearings used for inner race fault test.

Bearing type	Pitch diameter	Ball diameter	Number of balls	FCF
ER16K	38.52mm	7.94mm	9	$5.43f_r$

The collected vibration signal is shown in Fig. 5.11(a) and the measured shaft rotational speed (ISRF) is shown in Fig. 5.11(b). The shaft rotational frequency speeds up from 17.3Hz to 33Hz at 0-12s and then speeds down from 33Hz to 24.4Hz for the rest of the time span. Correspondingly, the amplitude of the vibration signal increases at the beginning and then reduces in Fig. 5.11(a). The TFR of the demodulated signal is obtained via the STFT shown in Fig. 5.11(c). Many ridges can be observed in the TFR, including the ISRF and the IFCF and its harmonics. The ridges appear to be cleaner than those in Fig. 5.10(c) for the outer race fault test.

According to the proposed procedure, the multiple T-F curve extraction algorithm is applied to the TFR. The extracted five T-F curves are shown in Fig. 5.11(d). The average ratios of the curves to the curve of the lowest curve are calculated as 1, 2.00, 5.43, 10.90, and 16.37, respectively. Among the calculated average ratios, 5.43 matches the FCC (5.43) perfectly. Additionally, 10.90 and 16.37 match 2*FCC and 3*FCC with relative error 0.4% and 0.5%, respectively. Also, 2.00 indicates a double of the curve of the lowest frequency. Therefore, according to the proposed fault diagnosis strategy, an inner race fault is detected. In addition, the extracted curve of the lowest frequency is regarded as the ISRF and the extracted curve of the average ratio 5.43 is regarded as the IFCF.

Comparisons of the extracted IFCF and the extracted ISRF to the measured curves are shown in Fig. 5.11(e). It can be seen that the extracted ISRF and the extracted IFCF well correspond to the measured ISRF and the calculated IFCF, respectively. Moreover, the relative error of the IFCF and the ISRF are calculated as 2.40% and 2.59%, respectively.

To compare the fast path-optimization approach with the maximum amplitude approach for T-F curve extraction, the extracted T-F curve by extracting the point of the maximum amplitude at each moment is shown in Fig. 5.11(f). Similar to the result shown in Fig. 5.10(f), the extracted curve is discontinuous with random jumps. Obviously, the proposed multiple T-F curve extraction algorithm based on the fast path-optimization approach is superior.

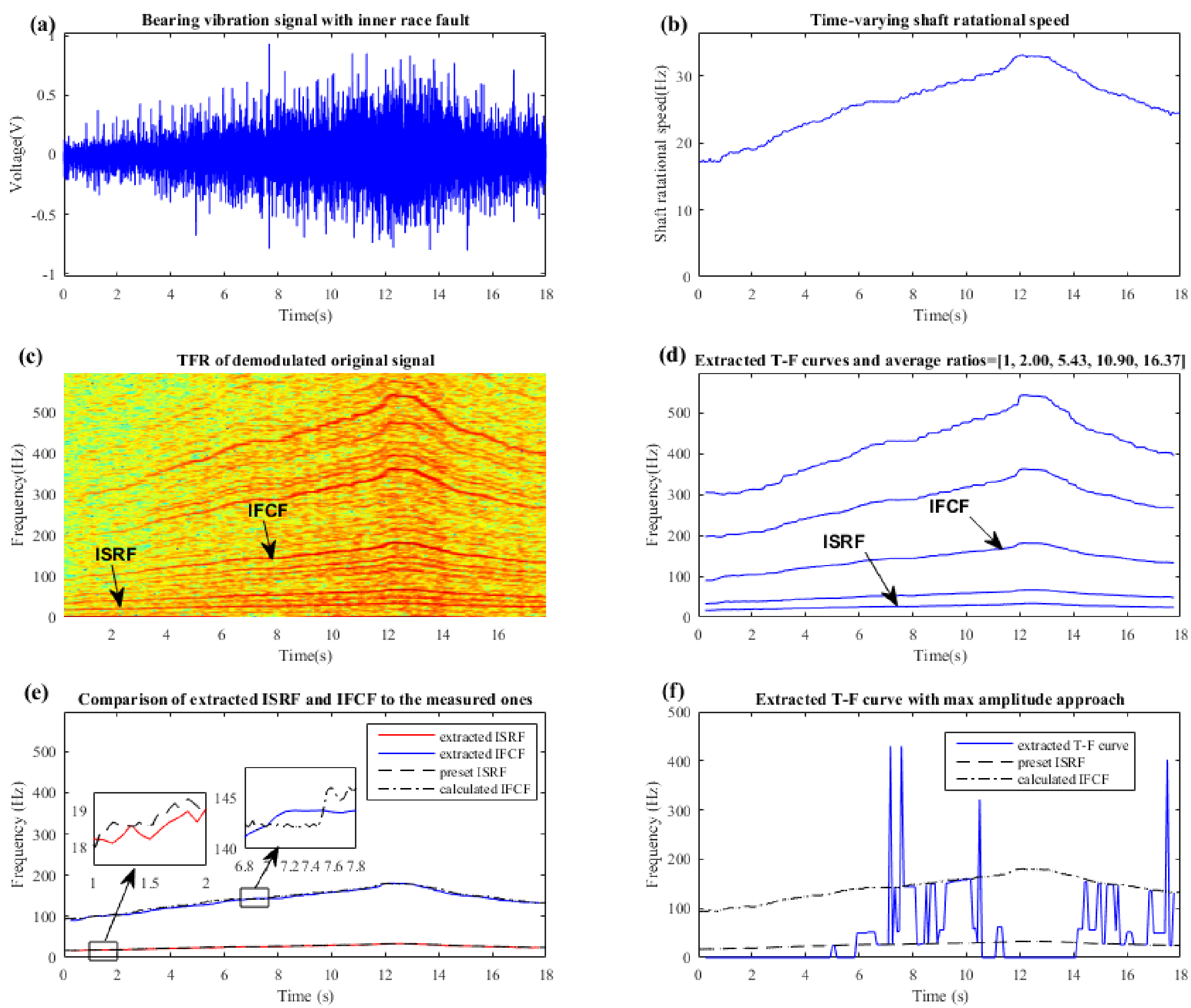

Fig. 5.11 Results of bearing with outer race fault. (a) Bearing vibration signal with inner race fault; (b) time-varying shaft rotational speed; (c) TFR of demodulated original signal; (d) extracted T-F curves; (e) comparison of extracted ISRF and IFCF to the measures curves; (f) extracted T-F curve with max amplitude approach.

5.6.3 Healthy bearing

To illustrate the applicability of the proposed bearing fault diagnosis method for healthy bearings, the left bearing shown in Fig. 5.9 is replaced by a healthy bearing. The parameters of the bearing used for the test are given in Table 5.7. For this type of bearing, the FCCs of outer race fault and inner race fault are calculated as 3.57 and 5.43, respectively. The sampling frequency of the signal is set as 12 kHz and the signal length is 8.65s.

The collected vibration signal is shown in Fig. 5.12(a) and the shaft rotational speed (i.e. ISRF) varies from 18.75 to 30.46 Hz with a pattern shown in Fig. 5.12(b). The amplitude of the vibration signal correspondingly increases slightly. The TFR of the demodulated signal is obtained via the STFT as shown in Fig. 5.12(c). In the TFR, no ridges can be clearly observed.

According to the proposed bearing fault diagnosis procedure, the multiple T-F curve extraction algorithm is applied to the TFR. The extracted five curves are plotted in the blue color shown in Fig. 5.12(d). The average ratios of the curves to the curve of the lowest frequency are calculated as 1, 1.47, 1.94, 2.25, and 3.83, respectively. Among these average ratios, none of them matches the FCC of the outer race fault (3.57) or the FCC of the outer race fault (5.43) within a 5% relative error. Therefore, according to the proposed fault diagnosis strategy, one more curve has to be extracted from the TFR beneath the extracted curve of the lowest frequency. The extracted curve is plotted in the red color shown in Fig. 5.12(d). Then, the average ratios of the six curves to the new curve are recalculated as 1, 1.58, 2.27, 3.04, 3.35, and 6.0, respectively. Similarly, none of the average ratios matches either the FCC of the outer race fault (3.57) or the FCC of the outer race fault (5.43) within a 5% relative error. Thus, according to the proposed procedure, the second extra curve has to be extracted from the TFR beneath curve plotted in red color. The second extra T-F curve is plotted in the black color shown in Fig. 5.12(d). The average ratios of the seven extracted curves to the new curve are recalculated as 1, 1.57, 2.50, 3.69, 4.84, 5.63, and 9.54, respectively. Again, these average ratios do not match either the FCC of the outer race fault (3.57) or the FCC of the outer race fault (5.43) within a 5% relative error. Therefore, the third extra curve has to be extracted from the TFR beneath the curve plotted in black color. The third extra curve is extracted and plotted with red dash line shown in Fig. 5.12(d). The corresponding average ratios of the eight curves to the new curve are calculated as 1, 2.42, 3.77, 5.94, 8.54, 11.45, 13.31, and 22.59, respectively. Once again, none of the calculated average ratios matches the FCC of the outer race fault (3.57) or the FCC of the outer race fault (5.43) within a 5% relative error. Since the maximum number of the extra curves was set as 3, the T-F curve extraction is terminated. Therefore, according to the proposed fault diagnosis strategy, the bearing is determined to be healthy.

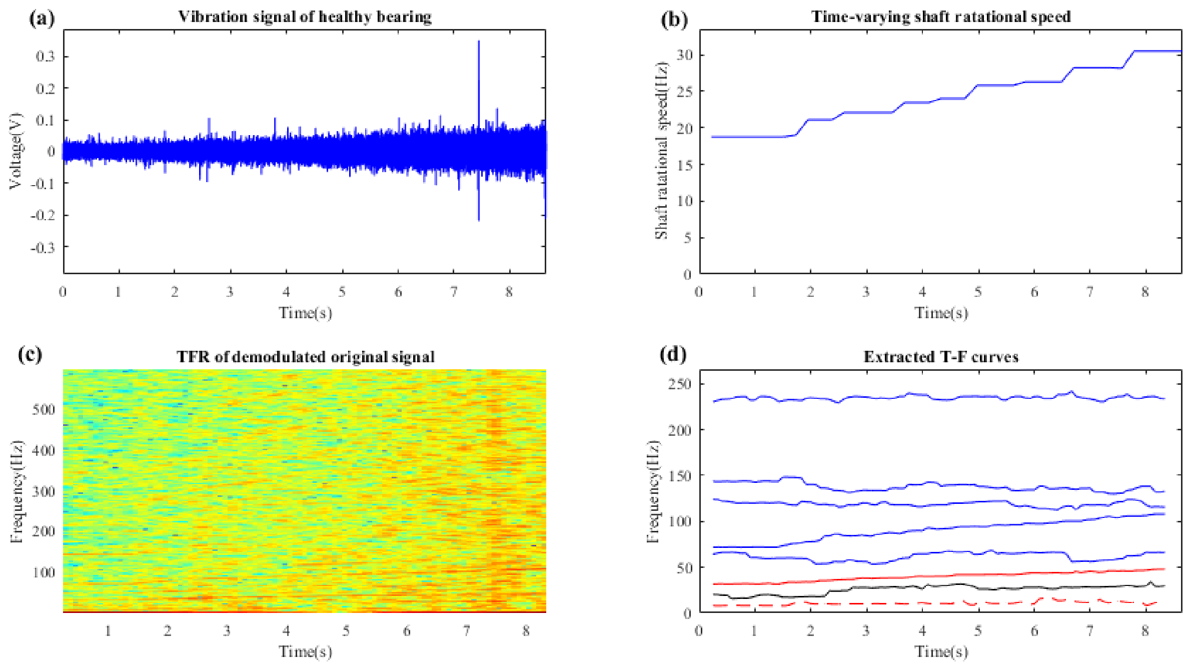

Fig. 5.12 Results of healthy bearing. (a) Bearing vibration signal; (b) time-varying shaft rotational speed; (c) TFR of demodulated original signal; (d) extracted T-F curves.

5.6.4 Data processing time

To examine the computational cost of this method in analyzing experimental data, the related processing times are listed in Table 5.8. Again, the method was performed using MATLAB R2015b, on the same computer used for simulations with a Windows 7 operating system, an Intel Core i5-4690 3.5 GHz processor and 8.0 GB of RAM. As shown in Table 5.8, the processing time required for the experimental data shows that the computational cost of the method is related to the sampling rate, signal length, and bearing health conditions.

Table 5.8 Processing time required for experimental data analysis.

	Sampling rate	Signal length	Number of extracted curves	Processing time	Average processing time for 1sec of data
Experiment 1	24kHz	5.3s	5	18.37s	3.47s
Experiment 2	12kHz	17.9s	5	31.81s	1.78s
Experiment 3	12kHz	8.65s	8	22.62s	2.62s

5.7 Conclusions

In this chapter, a multiple T-F curve extraction algorithm is proposed to extract T-F curves from the TFR for bearing fault diagnosis under time-varying speed conditions. Based on the multiple T-F curve extraction algorithm, a procedure for bearing fault diagnosis under time-varying speed conditions without tachometers is proposed.

The performance of the proposed procedure for bearing fault diagnosis under unknown time-varying speed conditions is tested with four types of simulated signals, including a signal with linearly increasing rotational speed which is dominant in the TFR, a signal with linearly increasing rotational speed which is not dominant in the TFR, a signal with an exponentially decaying rotational speed which is dominant in the TFR, and a signal with quadratically varying rotational speed which is dominant in the TFR. The results of simulations illustrate that the proposed procedure is effective regardless if the varying rotational speed is dominant or not in the TFR. At the same time, it also performed well for both linear and non-linear speed variation patterns. However, the tolerance for noise in the non-linearly varying rotational speed cases is lower than for the linearly varying rotational speed cases. Additionally, the results of the T-F curve extraction are influenced by the clearness of the ridges in the TFR.

The effectiveness of the proposed procedure is also validated with experimental data. The experimental data are collected from a bearing with an outer race fault, a bearing with an inner race fault and a healthy bearing, respectively. The results obtained by applying the proposed procedure to the experimental signals reveal that the proposed signal can be effectively used for bearing fault diagnosis under time-varying speed conditions without tachometers. For comparison, the result of the T-F curve extraction is also obtained by an existing method of extracting the maximum peak at each moment for each test. This clearly shows that the performance of the proposed T-F curve algorithm is superior.

It is worth mentioning that the results of the T-F curve extraction also depend on the resolution of the TFR. Advanced time-frequency analysis techniques developed based on continuous wavelet transform could potentially improve the accuracy of the T-F curve extraction. Additionally, the T-F curve extraction is not available for the extraction of curves that intersect each other. Therefore, for a bearing signal contaminated by interferences which may have frequency curves that intersect, the interferences must be removed from the signal before the T-F curve extraction can take place.

Chapter 6 Tachometer-free and resampling-free bearing fault diagnosis under time-varying speed conditions using OBSD and MTFCE

This chapter addresses objective 4. A tachometer-free and resampling-free method was proposed for bearing fault diagnosis under time-varying speed conditions with the presence of interference signals.

The content of this chapter has been submitted for publication.

6.1 Abstract

Bearing faults can be diagnosed in the frequency domain using the Fault Characteristic Frequency (FCF). However, the FCF becomes time-varying under time-varying speed conditions and hence cannot be employed directly. The common approach to time-varying speed requires a tachometer and signal resampling, which involves additional hardware and may lead to inaccurate results. In view of this, this chapter proposes a resampling-free and tachometer-free method for bearing fault diagnosis under time-varying speed conditions using a Time-Frequency approach. To address the effects of random noise and interferences, Oscillatory Behavior-based Signal Decomposition (OBSD) is used to extract bearing fault signatures from the contaminated signal by decomposing the signal according to oscillatory behavior instead of frequency. Subsequently, a Multiple Time-Frequency Curve Extraction (MTFCE) algorithm is implemented. With the proposed MTFCE algorithm, the Instantaneous Fault Characteristic Frequency (IFCF), its harmonics, and the Instantaneous Shaft Rotational Frequency (ISRF) can be extracted from the Time-Frequency Representation (TFR) and used for bearing fault diagnosis without signal resampling and tachometers. The proposed method of bearing fault diagnosis thus consists of four main steps: a) extract bearing fault signature from the contaminated signal via the OBSD, b) obtain the envelope of the extracted bearing fault signature via Hilbert transform and then convert the envelope into its TFR using the short-time Fourier transform, c) reveal the time-frequency curves from the TFR using the MTFCE, and d) identify the bearing fault type via the true ISRF following an ISRF-search diagnosis strategy. The effectiveness of the proposed method is validated using both simulated and experimental signals.

6.2 Introduction

Bearing fault diagnosis at constant rotational speed has been widely investigated ([7], [17], [58], [105], [9]). Vibration signal analysis is one of the most commonly used techniques for bearing fault diagnosis [17]. The local defect in a bearing induces an impulse response each time it strikes the contact surface. These impulse responses appear at a certain frequency called the Fault Characteristic Frequency (FCF) ([7],[58]). Each type of bearing fault has a specific FCF which is proportional to the rotational frequency [105]. The ratio of the FCF to the rotational frequency is determined by the structural parameters of the bearing [9], hence enabling bearing faults to be easily detected and diagnosed in the frequency domain. However, in reality rotating machines often operate under time-varying speed conditions which lead to a time-varying FCF. Additionally, the collected bearing signal is often contaminated by interferences transmitted from other resources, in addition to contamination by random noise. Therefore, there are two problems of interest to investigate for bearing fault diagnosis under time-varying speed conditions: bearing fault signature extraction from a contaminated signal under time-varying speed conditions and bearing fault diagnosis approaches for time-varying FCF.

For any fault diagnosis approach to work, it is critical to eliminate interferences and noise from the signal prior to the application of the fault diagnosis strategy. Many strategies for the elimination of interference and noise have been proposed in the literature. Band-pass filtering and reference based filtering are commonly used for bearing signal denoising and interference cancellation. Spectral Kurtosis (SK) is a useful technique developed for bearing signal denoising to determine the optimal band-pass filter since the bearing fault-induced signal has a relatively high resonance frequency ([23],[24]). However, the interference cannot be removed via spectral kurtosis if the interference has a high frequency or if the frequency of the interference is time-varying. Therefore, an adaptive filter which utilizes the Least Mean Square (LMS) algorithm has been employed to remove the interference for bearing fault diagnosis under time-varying speed conditions [29]. The filter can adaptively estimate the interference signal according to the reference signal correlated to the interference. By subtracting the estimated interference signal from the raw signal, the bearing fault signature can be obtained. However, a reference signal correlated to the interference is not always available. Therefore, frequency-independent and reference-free methods for bearing fault signature extraction under time-varying speed conditions are of interest to explore.

Oscillatory Behavior-based Signal Decomposition (OBSD) is a new technique which decomposes a signal according to oscillatory behavior [40] instead of frequency and does so without reference signals. The OBSD can be used to effectively extract the bearing fault signature from a signal blurred by interferences since the bearing fault signature can be considered as low oscillatory and the interference can be considered as high oscillatory[67].

An iterative OBSD method has been developed to remove multiple sets of interferences for bearing fault diagnosis [69]. Additionally, an auto-OBSD method was developed to automatically select parameters related to the OBSD for bearing fault signature extraction [106]. However, to date the OBSD has been only applied for bearing signature extraction under constant speed conditions. Theoretically, the OBSD method should still be effective under time-varying speed conditions since the oscillatory behavior of the bearing fault signature and the interference are not changed. Therefore, in this chapter, the OBSD is employed to extract bearing fault signature extraction under time-varying speed conditions and its effectiveness is investigated.

For bearing fault diagnosis under time-varying FCF, order tracking is one of the conventional methods used with time-varying speed conditions. By sampling the signal at constant angular (order) increments, instead of constant time increments, the time varying FCF is converted into constant order features [107]. With the resampled signal, fault diagnosis can then be achieved in the order domain according to the fault characteristic order [73]. The implementation of order tracking requires extra instruments, such as tachometers, and also additional computational operations [33], such as signal interpolation. However, tachometers are not always available or feasible to be installed on every machine of interest, and additionally the accuracy of order tracking is limited by interpolation [34]. Therefore, it is important to develop tachometer-free and resampling-free methods for bearing fault diagnosis under time-varying speed conditions.

Tachometer-free order tracking methods for bearing fault diagnosis under time-varying speed conditions have been investigated based on time-frequency analysis. With the development of time-frequency analysis techniques, such as the Short-Time Fourier Transform (STFT) and wavelet transform, the instantaneous frequency can be shown in a Time-Frequency Representation (TFR) [70]. For a bearing fault induced signal, the Instantaneous Fault Characteristic Frequency (IFCF), IFCF harmonics, Instantaneous Shaft Rotational Frequency (ISRF) and its harmonics can be observed as Time-Frequency (T-F) curves, also called ridges, in the TFR. In [108], the ISRF was estimated by searching the point which has the local maximum amplitude at each instant. With the estimated ISRF, the signal can be resampled without tachometer measurement of the shaft rotational speed. However, the estimated ISRF may not be accurate via the local maximum amplitude method if the signal is blurred by noise. Therefore, a cost function based method was employed to extract the ISRF for tachometer-free order tracking [109]. The cost function based method was related to both the amplitude of the point and the frequency difference between two connected points, which proved to be more reliable than the local maximum amplitude. In addition, to improve the local maximum method, an amplitude sum-based method was proposed to extract the IFCF from the TFR since the point with the maximum amplitude could be the IFCF or any of its harmonics [32]. All these methods for instantaneous frequency extraction are based on the assumption that the extracted T-F curve is known to be a certain harmonic of the ISRF or the IFCF. However, the extracted T-F curve could be any of them. Therefore, it is necessary to develop a more reliable method for T-F curve extraction without the above assumption. Moreover, these tachometer-free order tracking methods still rely on signal resampling.

Iterative Generalized Demodulation (IGD) based methods have been proposed for bearing fault diagnosis which are both tachometer-free and resampling-free. The generalized demodulation can be used to convert a curve in the TFR of a signal into a straight line [110]. By iterative application of generalized demodulation, an order spectrum can be obtained for fault diagnosis without signal resampling [99]. Following this idea, the IFCF was estimated by taking the point of maximum amplitude at each instant in the TFR, and the IFCF was straightened via generalized demodulation, then harmonics of the IFCF were straightened iteratively [81]. However, similar to the problem present in the tachometer-free order tracking methods, the extracted T-F curve may not be the IFCF as assumed. Therefore, a reliable T-F curve extraction algorithm is critical in any tachometer-free and resampling-free methods for bearing fault diagnosis under time varying speed conditions. Moreover, if the ISRF, the IFCF, and their harmonics can be extracted directly from the TFR, then bearing fault diagnosis can be completed without using generalized demodulation.

In this chapter, a resampling-free and tachometer-free method is proposed for bearing fault diagnosis under time-varying rotational speed conditions. The OBSD, which is frequency independent and reference-free, is employed to extract the bearing fault signatures from the contaminated signal. Subsequently, a Multiple Time-Frequency Curve Extraction (MTFCE) algorithm is implemented to extract multiple T-F curves directly from the TFR of the decomposed signal without using generalized demodulation. The MTFCE algorithm iteratively applies path optimization to extract the T-F curves. Finally, a bearing fault diagnosis strategy is proposed to utilize the extracted T-F curves for fault diagnosis. The effectiveness of the proposed method is validated by simulations and experiments. The remainder of the chapter is structured as follows. In section 6.3, the proposed bearing fault diagnosis method is presented. Simulations are then conducted to examine the effectiveness of the proposed method in section 6.4. The simulated signals cover both conditions where the IFCF and the Instantaneous Interference Frequency (IIF) cross and do not cross each other. In section 6.5, signals collected from experiments are used to further test the effectiveness of the proposed method. Finally, in section 6.6, conclusions are drawn based on the results in section 6.4 and section 6.5.

6.3 Proposed method for bearing fault diagnosis under unknow time-varying speed conditions

A schematic of the proposed method for bearing fault diagnosis under unknown time-varying speed conditions is shown in Fig. 6.1. The proposed method consists of 4 main steps. In step 1, the OBSD method is applied to the raw signal which is contaminated by interferences and noise. The raw signal is decomposed into a low oscillatory component, a high oscillatory component and the residual by the OBSD. The low oscillatory component is considered as the bearing fault signature and the high oscillatory component is considered as the interference due to features of the oscillatory behavior. Step 2 involves the use of the Hilbert transform and the STFT to obtain the envelope TFR of the extracted bearing fault signature. The bearing fault-induced signal can be regarded as an amplitude demodulated signal with the excited resonance frequency as the frequency of the carrier signal; hence by applying the Hilbert transform, the excited resonance is demodulated and the obtained signal is called the envelope [9]. The STFT is one of the most commonly used techniques to obtain the TFR of a signal with time-varying frequency. For a signal $x(t)$, the TFR of a signal $x(t)$ is obtained via STFT as

$$X(\tau, f) = \int_{-\infty}^{+\infty} x(t) w(t-\tau) e^{-j2\pi ft} dt \tag{6.1}$$

where $w(t)$ is the window function, t is the time variable, and f is the frequency variable. The IFCF and its harmonics should show ridges in the envelope TFR of the low oscillatory component. To simplify the description, the TFR of the envelope of a signal is called the envelope TFR of a signal in this chapter. In step 3, the MTFCE algorithm is employed to extract T-F curves from the TFR obtained in step 2. The extracted T-F curves are labeled as T-F curves (L) in the diagram shown in Fig. 6.1. The IFCF and its harmonics are expected to be extracted from the TFR. In step 4, a fault diagnosis strategy is proposed to diagnose the condition of the bearing. If the ISRF also appears as a clear ridge in the TFR obtained in step 2 and extracted by the MTFCE, the fault type of a faulty bearing can be determined by the ratio of the IFCF to the ISRF, called Fault Characteristic Coefficient (FCC). However, the ISRF may not be present in the TFR obtained in step 2. This problem is also considered and solved by the proposed fault diagnosis strategy. The OBSD, MTFCE, and the ISRF-search diagnosis strategy are detailed in the following subsections.

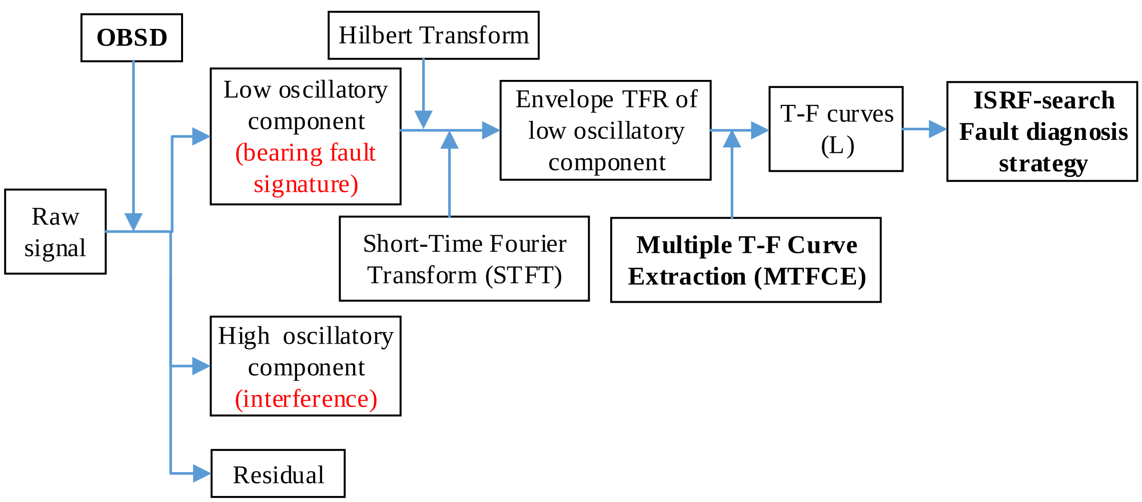

Fig. 6.1 Procedure of the proposed bearing fault diagnosis under time-varying speed conditions

6.3.1 OBSD for bearing fault signature extraction

The essence of the OBSD method is using two set of wavelets with two different oscillatory behaviors to estimate a given signal [40]. The signal is decomposed into a low oscillatory component and a high oscillatory component by the OBSD. According to the features of the bearing fault-induced signal and the interference signal, the bearing fault-induced signal is more impulsive which can be considered as low oscillatory behavior, and the interference is smoother which can be considered as high oscillatory behavior [67]. Therefore, the OBSD method can be used to extract the bearing fault signature from a signal obscured by interferences.

Compared to frequency or scale based methods for interference removal, the OBSD is superior since the signal decomposition is based on oscillatory behaviors instead of frequencies. A Q-factor defined as the ratio of the center frequency to the bandwidth of the frequency response is used to describe the oscillatory behavior of a wavelet [62]. A higher value of the Q-factor indicates a higher level of oscillation. In Fig. 6.2, four wavelets and their frequency spectra are shown. The Q-factors of the waveforms are also given. It can be seen that waveform 1 and waveform 2 have different oscillatory behaviors, however, their frequency spectra share the same center frequency. The same can be observed for waveform 3 and waveform 4. Under such circumstances, frequency based band-pass filters would not be able to separate waveforms 1 and 2, nor waveforms 3 and 4. However, they can be separated according to their Q-factors. It is calculated that waveforms 1 and 3 have a Q-factor of 1 since they exhibit low oscillatory behavior, and waveforms 2 and 4 have a Q-factor of 5 since they have relatively high oscillatory behavior. Additionally, this demonstrates that the waveforms that have the same oscillatory behaviors have the same Q-factors, even if the center frequencies are different. This makes the OBSD effective in capturing the true features that are useful for bearing fault signature extraction under time-varying speed conditions.

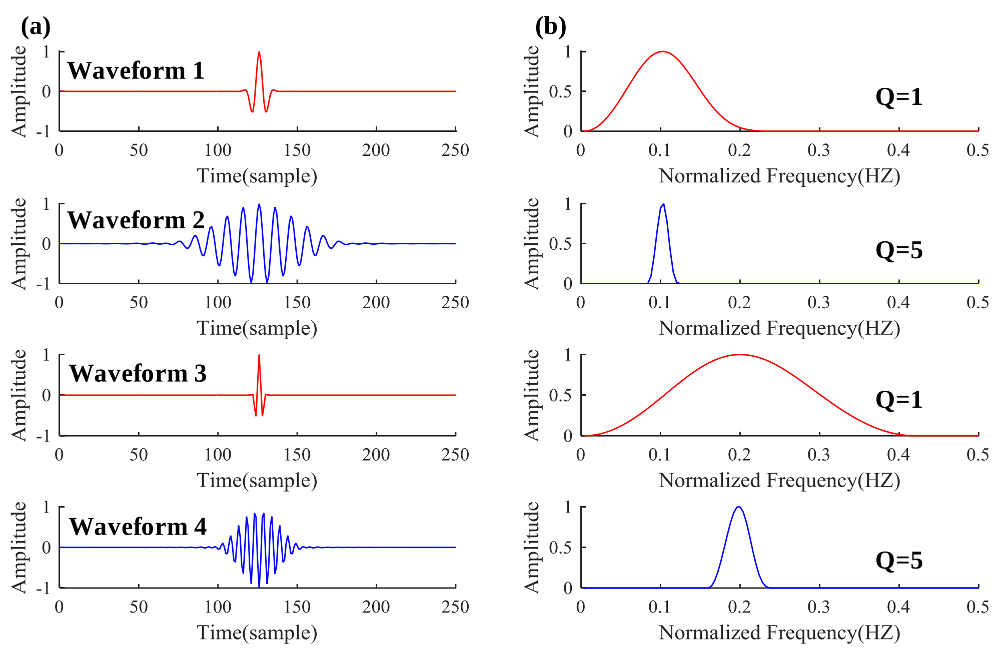

Fig. 6.2 Wavelets and their frequency spectra (cited from [62] with minor changes)

The OBSD method employs the Tunable Q-factor Wavelet Transform (TQWT) and Morphological Component Analysis (MCA) to realize the signal decomposition [40]. The TQWT is used to generate a set of wavelets that have the same Q-factor, i.e. the same oscillatory behavior. The wavelets can be obtained with the selection of three parameters, Q (Q-factor), r and P, where Q is related to the oscillatory behavior, r is related to the redundancy of the frequency responses, and P is the number of wavelets. In Fig. 6.3, a set of wavelets with Q=3, r=3, and P=15 and the corresponding frequency spectra are shown. By setting up two sets of wavelets, one set hainv low oscillatory behavior with parameters Q_l, r_l, and P_l, and the other set haing high oscillatory behavior with parameters Q_h, r_h, and P_h, the signal decomposition can be completed by MCA. The wavelet coefficients are obtained via optimization [40] as

$$\begin{bmatrix} \mathbf{w}_h^{opt} \\ \mathbf{w}_h^{opt} \end{bmatrix} = \underset{\mathbf{w}_l,\mathbf{w}_h}{arg\min} \left\| \mathbf{y} - \mathbf{S}_l \mathbf{w}_l - \mathbf{S}_h \mathbf{w}_h \right\|_2^2 + \lambda_l \left\| \mathbf{w}_l \right\|_1 + \lambda_h \left\| \mathbf{w}_h \right\|_1 \tag{6.2}$$

where $\mathbf{w}_l$ refers to wavelet coefficients for the low oscillatory component, $\mathbf{w}_h$ stands for the wavelet coefficients for the high oscillatory component, $\mathbf{w}_l^{opt}$ and $\mathbf{w}_h^{opt}$ are results after the optimization, $\mathbf{y}$ is the signal to be decomposed, $\mathbf{S}_l$ refers to wavelets obtained via the TWQT for the low oscillatory component, $\mathbf{S}_h$ represents wavelets for the high oscillatory component, λ_l is the regularization parameter for the low oscillatory component, λ_h is the regularization parameter for the high oscillatory component, $\| \|_1$ and $\| \|_2$.are the norm-1 operation and the norm-2 operation, respectively. This optimization problem can be solved using an iterative algorithm called Split Augmented Lagrangian Shrinkage Algorithm (SALSA), obtained as [40]

$$\mathbf{w}_i^k = f\left(\mathbf{w}_i^{k-1}, \mu\right),\ i = l, h,\ k = 1, 2, ..., K \tag{6.3}$$

where μ is the penalty parameter and K is the maximum number of iterations. Details of the solution can be found in appendix A. By selecting the maximum number of iterations, the optimal wavelet coefficients are obtained as

$$\mathbf{w}_i^{opt} = \mathbf{w}_i^K,\ i = l, h \tag{6.4}$$

With the calculated wavelet coefficients, the decomposed low oscillatory and high oscillatory components can then be obtained by inverse TQWT with the optimized wavelet coefficients.

There are 10 parameters related to the OBSD method: Q_l, r_l, P_l, Q_h, r_h, P_h, λ_l, λ_h, μ, and K. These parameters have to be appropriately selected for the application of the OBSD. The Q-factor is related to the oscillatory level of the target signal. For bearing fault signature extraction, Q_l can be selected as 1 since the bearing fault-induced signal is considered impulsive. Q_h should be properly selected to match the oscillatory level of the interferences. The increase of the redundancy r can improve the density of the frequency responses, but also increases the computational cost. In this research, it is found that $r_l = r_h$ =6 can be a proper choice to balance the accuracy of the result and the computational cost. Additionally, to ensure the accuracy, the number of wavelets P should always be selected as the maximum number which is determined by Q and r. The selection of regularization parameters λ_l and λ_h is a trade-off between small residual and high sparsity of the obtained wavelet coefficients. Hence they should be properly selected to balance the residual and the sparsity. Here λ_l is set equal to λ_h. The penalty factor μ is related to the convergence speed of the optimization solution. It should be selected as a value such that fast convergence can be achieved. The maximum number of iterations K should be selected large enough to ensure the acceptable level of convergence at acceptable computational cost. A more detailed guidance on the parameter selection can be found in [106].

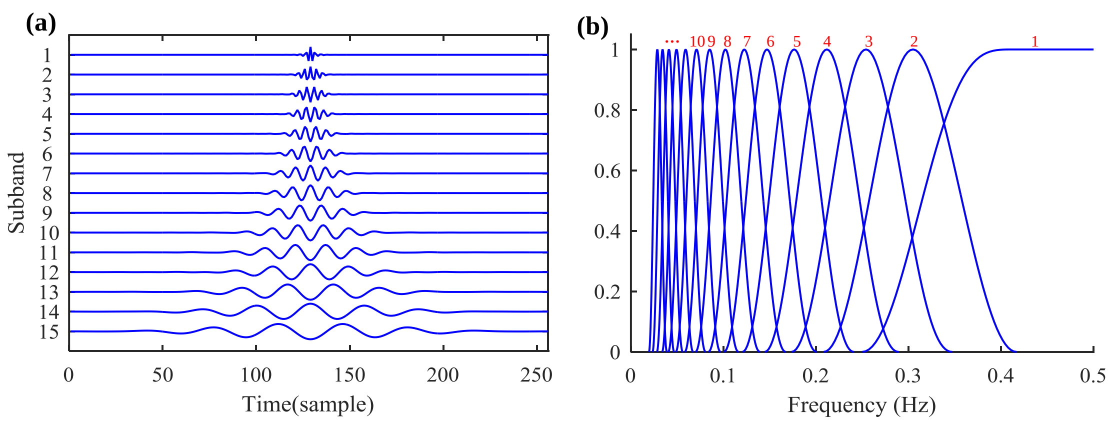

Fig. 6.3 Wavelets of each sub-band and the corresponding frequency response of TQWT when Q=3, r=3, and P=15

For bearing fault signature extraction from a signal contaminated by interference and noise, the bearing fault-induced signal is impulsive, which is low oscillatory, and the interference signal is smoother, which is high oscillatory. Therefore, the decomposed low oscillatory component and the high oscillatory component via the OBSD are taken to be the bearing fault signature and interference, respectively.

6.3.2 Multiple Time-Frequency Curve Extraction (MTFCE) algorithm

An algorithm for MTFCE was proposed for bearing fault diagnosis under time-varying speed conditions in [111]. With this algorithm, multiple T-F curves can be extracted from the TFR of a signal. The algorithm iteratively utilizes fast path optimization for T-F curve extraction.

Fast path optimization for T-F curve extraction from the TFR is achieved by optimally extracting a string of peaks along the time span [103]. A scheme of the T-F curve extraction is shown in Fig. 6.4. The circles in the TFR represent peaks at each moment τ_n. The number of peaks at τ_n is denoted as $N_p(\tau_n)$. For the *m*th peak at τ_n, the amplitude is denoted as $Q_m(\tau_n)$ and the frequency is denoted as $v_m(\tau_n)$, respectively. A peak map is constructed with all these peaks. Each peak at τ_n links one peak at the past moment τ_{n-1}. The linked peak at τ_{n-1} can be determined by optimization with a cost function

$$F[\ ]=\begin{cases} \log Q_m(\tau_n) & n=1 \\ \log Q_m(\tau_n)+w_2\left(v_m(\tau_n),\mathrm{m}[f_d],\mathrm{IQR}[f_d],\lambda_2\right) & n=2 \\ \log Q_m(\tau_n)+w_2\left(v_m(\tau_n),\mathrm{m}[f_d],\mathrm{IQR}[f_d],\lambda_2\right) & \\ +w_1\left(v_m(\tau_n)-f_d(\tau_{n-1}),\mathrm{m}[\Delta f_d],\mathrm{IQR}[\Delta f_d],\lambda_1\right) & n\geq 3 \end{cases} \tag{6.5}$$

Where

$$w_1\left(v_m(\tau_n)-f_d(\tau_{n-1}),m[\Delta f_d],IQR[\Delta f_d],\lambda_1\right)=-\rho_1\left|\frac{v_{m_n}(\tau_n)-f_d(\tau_{n-1})-m[\Delta f_d]}{IQR[\Delta f_d]}\right| \tag{6.6}$$

$$w_2\left(v_m(\tau_n),m[f_d],IQR[f_d],\lambda_2\right)=-\rho_2\left|\frac{v_m(\tau_n)-m[f_d]}{IQR[f_d]}\right| \tag{6.7}$$

$$m[\ ]=\mathrm{perc}_{0.5}[],\quad IQR=\mathrm{perc}_{0.75}[]-\mathrm{perc}_{0.25}[] \tag{6.8}$$

and where $f_d\left(\tau_{n-1}\right)$ is the frequency of the candidate peak at τ_{n-1}, f_d are the frequencies of a series of candidate ridge points in history $[\tau_1,...,\tau_{n-1}]$, Δf_d is the derivative of f_d, $\mathrm{perc}_p[f(t)]$ is the pth quantile of $f(t)$, ρ_1 and ρ_2 are penalty factors which can be taken as 1 according to the analysis in [103], and $w_1()$ and $w_2()$ are weight functions used to suppress the atypical variations of the ridge frequency value and derivative, respectively. After every peak links to one of the peaks at the previous moment, many paths are generated in the peak map, as shown in Fig. 6.4(b). Finally, by selecting the point of the minimum value of the cost function at the last moment, the whole T-F curve can be extracted following the path as shown in Fig. 6.4(c). Compared to T-F curve extraction methods by searching local maxima, the fast path optimization method can effectively prevent unexpected frequency jumps in the extracted T-F curve [103].

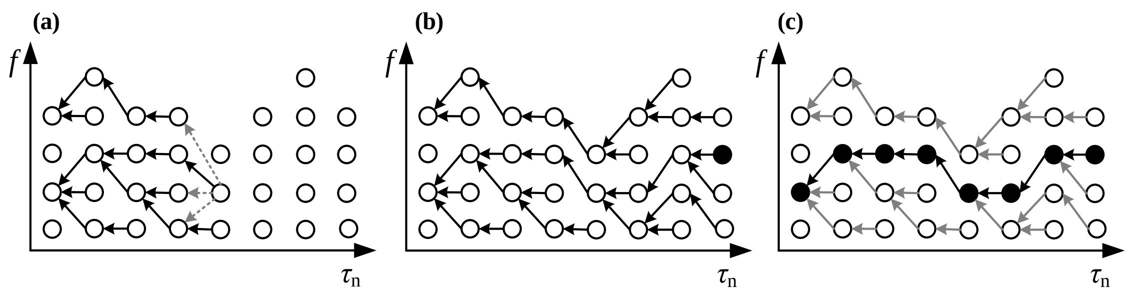

Fig. 6.4 Scheme of T-F curve extraction from the TFR

The MTFCE algorithm is developed by iteratively utilizing the fast path optimization method. The procedure of the algorithm is as follows:

Step 1. Extract one T-F curve from the TFR using the fast path optimization method;

Step 2. Remove the peaks of the last extracted T-F curve from the peak map and generate a new peak map;

Step 3. Extract a new T-F curve from the new peak map via the fast path optimization method;

Step 4. Repeat steps 2 and 3 to extract a new T-F curve until the number of extracted curves reaches the preset maximum number $C_{\max}$.

The maximum number of extracted T-F curves should ensure that enough curves are extracted from the TFR. However, the computational cost increases with the increase of the number of extracted curves. Therefore, $C_{\max}$ should be selected as large enough within the acceptable computational cost. In this chapter, $C_{\max}$ is set as 4 in sections 6.4 and 6.5.

Applying the MTFCE algorithm to the envelope TFR of the low oscillatory component, the IFCF of the bearing and its harmonics can be extracted if the bearing is faulty. For bearing fault diagnosis, the ISRF also has to be extracted from the TFR. If the bearing fault-induced signal is amplitude modulated with the ISRF as the frequency of the modulation signal, then the ISRF will also appear in the envelope TFR of the low oscillatory component and can be extracted via the MTFCE algorithm. However, the ISRF may not appear in the envelope TFR of the low oscillatory component if the bearing fault-induced signal is slightly or not amplitude modulated with the ISRF as the frequency of modulation. This is challenging for tachometer-free bearing fault diagnosis under time-varying speed and will be addressed in the following subsection.

6.3.3 An ISRF-search fault diagnosis strategy

With the output of the OBSD and MTFCE algorithm, a bearing fault diagnosis strategy is proposed for fault diagnosis under unknown time-varying speed conditions. It is known that the FCF for each fault type is proportional to the shaft rotational speed and the ratio of the FCF to the rational frequency is constant [9]. In this chapter, this ratio is called the Fault Characteristics Coefficient (FCC). For bearings with outer race fault or inner race fault, the FCF and FCC can be calculated by

$$\text{BPFO}=\frac{n_b}{2}\left(1-\frac{d}{D}\cos\phi\right)f_r = FCC_O \cdot f_r \tag{6.9}$$

$$\text{BPFI}=\frac{n_b}{2}\left(1+\frac{d}{D}\cos\phi\right)f_r = FCC_I \cdot f_r \tag{6.10}$$

where n_b is the number of rolling elements, d is the diameter of the rolling element, D is the pitch diameter of the bearing, ϕ is the angle of the load from the radial plane, f_r is the shaft rotational frequency, FCC_O and FCC_I are the outer race FCC and inner race FCC, respectively. Since the FCC is independent of the shaft rotational frequency, bearing faults can still be diagnosed by the FCC under time-varying rotational speed conditions. Therefore, if both the IFCF and the ISRF are extracted, the fault can be diagnosed by matching the average ratio of IFCF/ISRF to the FCC. The average ratio is defined as

$$R_a = \frac{1}{N}\sum_{n=1}^{N}\frac{f_p\left(\tau_n\right)\_i}{f_p\left(\tau_n\right)\_l} \tag{6.11}$$

where N is the length of the extracted curves, $f_p\left(\tau_n\right)\_l$ denotes the bottom T-F curve, i.e. the T-F curve with the lowest frequency, and $f_p\left(\tau_n\right)\_i$ is one of the extracted T-F curves. The calculated average ratio may not exactly match the FCC due to the limited resolution of the TFR obtained via the STFT. Therefore, a 5% relative error of the calculated average ratio to the FCC is proposed for the tolerance of matching, i.e. $\left|R_a - \mathrm{FCC}\right| \bullet 100\% / \mathrm{FCC} \le 5\%$. However, the ISRF may not be extracted from the envelope TFR of the low oscillatory component since the ISRF may not appear in the TFR. Typically, the envelope of the inner race fault-induced signal is amplitude modulated with the shaft rotational frequency as the modulation frequency, however, the envelope of the outer race fault-induced signal is not amplitude modulated [9]. Therefore, theoretically, the ISRF appears in the envelope TFR of the extracted bearing fault signature caused by an inner race fault, but is absent in the envelope TFR of the extracted bearing fault signature caused by an outer race fault. Considering the fact that the shaft rotational frequency is always contained in the collected bearing vibration signal, it is possible that the ISRF appears in the TFR of the high oscillatory component as one of the interferences, if it does not appear in the envelope TFR of the low oscillatory component. Actually, the vibration signal often contains the interference with the ISRF as the fundamental frequency caused by misalignment, eccentricity or imbalance [109]. Therefore, the TFR of the decomposed high oscillatory component is also important for bearing fault diagnosis. Based on this, a bearing fault diagnosis strategy can be carried out to address the problem of the ISRF not being extracted from the envelope TFR of the low oscillatory component.

The flowchart of the proposed fault diagnosis strategy is shown in Fig. 6.5. If the ISRF is included in the extracted T-F curves from the envelope TFR of the low oscillatory component, the ISRF should be the curve at the bottom since the ISRF is smaller than the IFCF and its harmonics. By matching the calculated average ratios to the FCCs of the bearing, the bearing fault type can be determined if the bearing is faulty. However, if the ISRF is not extracted from the envelope TFR of the low oscillatory component, the TFR of the decomposed high oscillatory component has to be further investigated. Therefore, if none of the calculated average ratios for the extracted T-F curves (L) match the FCC, the STFT is applied to the decomposed high oscillatory component to obtain the TFR of the high oscillatory component. Then, the MTFCE algorithm is employed again to extract T-F curves (H) from the TFR of the high oscillatory component. Similarly, the extracted curve with the lowest frequency is considered as the ISRF. By taking the new bottom curve as the curve with the lowest frequency among the extracted T-F curves (H), the average ratios for the extracted T-F curves (L) are re-calculated. If the re-calculated average ratios match one of the FCC and at least one of its harmonics, then the fault type can be determined. Otherwise, the bearing is considered as healthy.

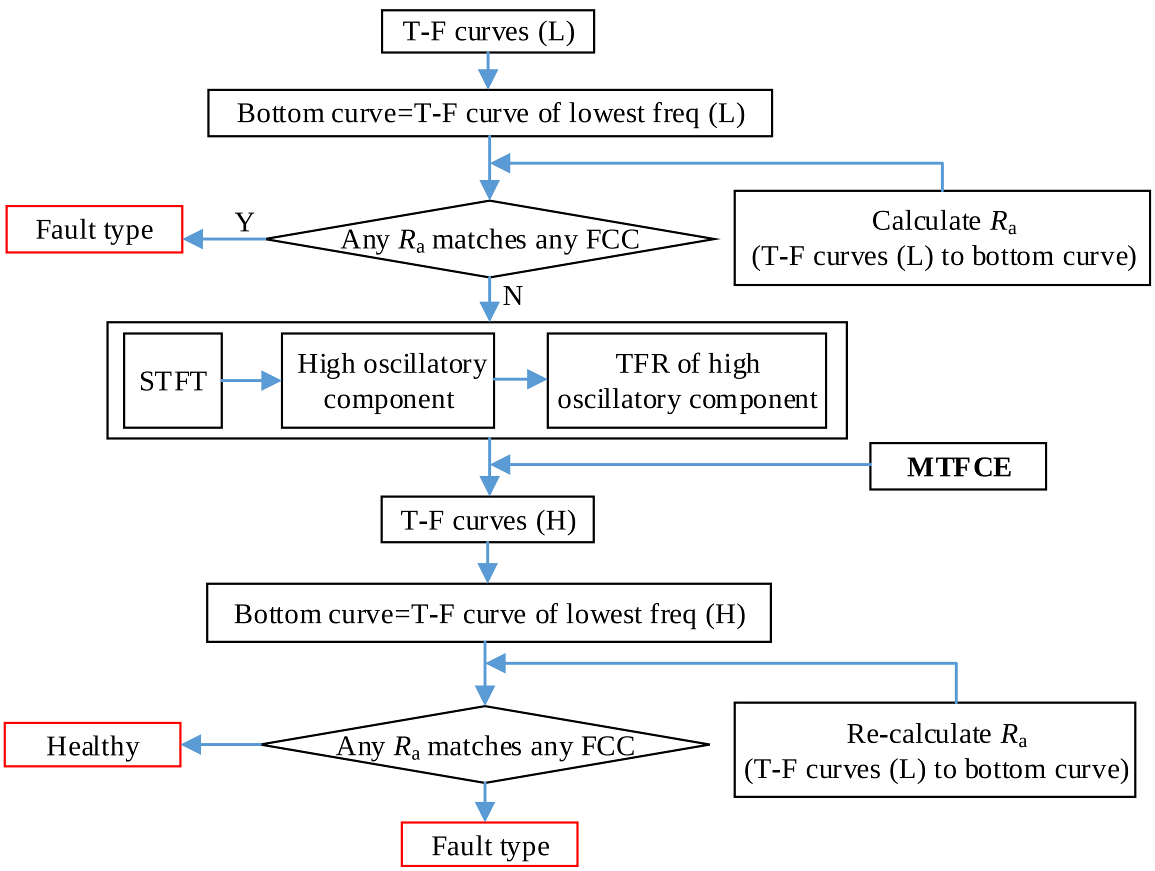

Fig. 6.5 Flowchart of the ISRF-search fault diagnosis strategy

6.4 Simulations

To investigate the effectiveness of the proposed method for bearing fault diagnosis under time-varying speed conditions, simulations are conducted. The signal is simulated as a mixture of bearing fault-induced signal, interferences and random noise.

The bearing fault-induced signal can be simulated as impulse responses which occur at the FCF along the time span [32]. For a bearing operating under time-varying speed conditions, the equation is given as [81]

$$\begin{aligned} x_B(t) &= \left[1+A(t)\right]\sum_{m=1}^{M} s_m(t) \\ &= \left[1+\alpha\cos(2\pi f_r)\right]\sum_{m=1}^{M} L_m e^{-\beta(t-t_m)} \sin\left[\omega_r(t-t_m)+\phi_m\right] u(t-t_m) \end{aligned} \tag{6.12}$$

where $A(t)=\alpha\cos(2\pi f_r t)$ represents the modulated waveform of frequency f_r, α is the amplitude of the modulation ($\alpha<1$), M is the number of impulse responses which is determined by the signal length T and FCF, $s_m(t)$ represents the *m*th impulse response, L_m is the amplitude of the *m*th impulse response, β is the coefficient related to damping, ω_r is the excited resonance frequency or damped frequency of the vibration system, ϕ_m is the phase of the *m*th impulse response, and $u(t)$ is unit step function. In the previous equation, t_m is the occurrence time of the *m*th impulse response which is calculated as

$$\begin{cases} t_1 = (1+\delta_1)\left[1/f_c(0)\right] \\ t_m = (1+\delta_m)\left[1/f_c(0)+1/f_c(t_1)+\ldots+1/f_c(t_{m-1})\right] \quad m=2,3,\ldots,M \end{cases} \tag{6.13}$$

where δ_m is the random slippage ratio with the average varying between 0.01 and 0.02, $f_c(t)$ denotes the IFCF calculated by $f_c(t)=FCC * f_r(t)$, and the time interval between the (*m*-1)th impulse response and the *m*th impulse response is $(1+\delta)/f_c(t_{m-1})$.

One set of the interference can be simulated as the sum of sinusoidal functions of the frequency of the interference and its harmonics, given as [81]

$$I(t)=\sum_{n_i=1}^{N_i} B_{n_i}\sin(2\pi n_i f_I t) \tag{6.14}$$

where N_i is the number of sinusoidal functions, B_{n_i} is the amplitude, and f_I is the time-varying frequency of the interference, called Instantaneous Interference Frequency (IIF) in this chapter. The interference contained in the signal is possibly made of more than one set. It is indicated in [109] that the vibration signal also comprises the interference with the ISRF as the fundamental frequency caused by misalignment, eccentric or imbalance. Therefore, the interference signal can be simulated as

$$x_I(t) = I_1(t) + I_2(t) = \sum_{n_{1i}=1}^{N_1} B_{1n_i} \sin(2\pi n_{1i} f_I t) + \sum_{n_{2i}=1}^{N_2} B_{2n_{2i}} \sin(2\pi n_{2i} f_r t) \tag{6.15}$$

Then, the faulty bearing signal contaminated by interferences and noise is simulated by

$$x(t) = x_B(t) + x_I(t) + n(t) \tag{6.16}$$

where $n(t)$ represents random noise with intensity measured by the SNR.

An example of the simulated bearing fault-induced signal and an example of interference signal are shown in Fig. 6.6(a) and Fig. 6.6(b), respectively. The bearing fault-induced signal is simulated by Eq. (6.12) with f_r =20t+10 Hz, α =0.95, L_m =1, β =1500, ω_r =4000*2π rad/s, ϕ_m =0, δ_m is a random number with a mean of 0.01, FCC=3.7, and T=1s which determines M =74. The interference is simulated by Eq. (6.15) with N_I =1, N_i =3, f_I =50t+20 Hz, and B =[1, 0.3, 0.1]. As shown in Fig. 6.6(a), the bearing fault-induced signal is composed of impulse responses which die out with an exponential decay; this can be considered as low oscillatory behavior. Also, the maximum amplitudes of the impulse responses are varied in proportion to the waveform $A(t)$ since the signal is amplitude modulated with $A(t)$ as the modulation waveform. The interval between two responses is the reciprocal of the IFCF. It can be seen that the interval becomes smaller with an increase in the shaft rotational frequency f_r . Compared to the bearing fault-induced signal, the interference signal shown in Fig. 6.6(b) is smoother since the signal is continuously oscillating. Therefore, the interference can be considered as high oscillatory behavior. It can be seen that even if the frequencies are time-varying, the bearing fault-induced signal and the interference signal can still be regarded as low oscillatory and high oscillatory, respectively.

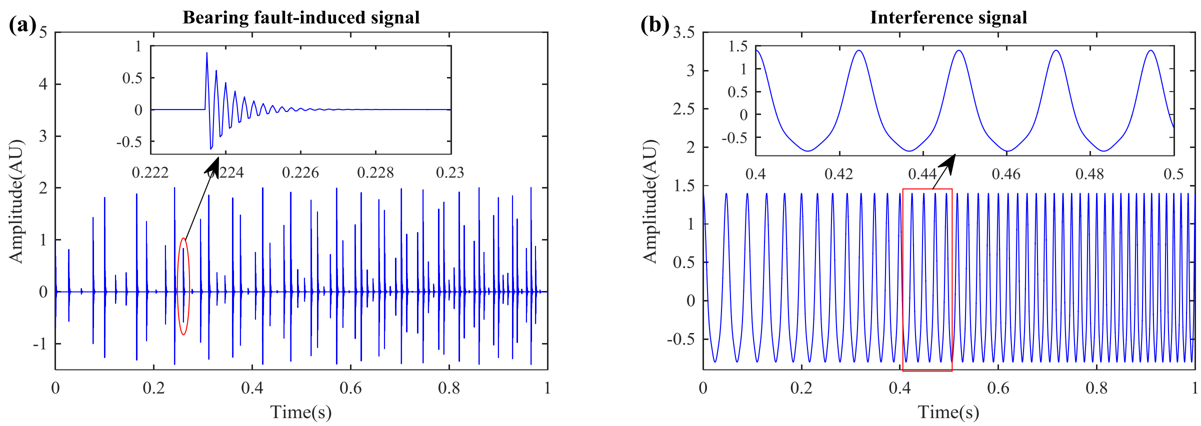

Fig. 6.6 Examples of the simulated bearing fault-induced signal and the interference signal. (a) Simulated bearing fault-induced signal, and (b) simulated interference signal.

6.4.1 Cases when IFCF and IIF do not cross

The simulation starts with a simulated signal in which the IFCF does not cross the IIF, i.e. the IFCF and the IIF are proportional to each other. To test the effectiveness of the OBSD for bearing fault signature extraction, the IFCF and the IIF are chosen to be very close.

6.4.1.1 Case 1: ISRF is not modulated by the bearing fault-induced signal

In this simulation, the raw signal is composed of a bearing fault-induced signal, interference signal, and random noise. To simulate the signal of faulty bearing with outer race fault, the bearing fault-induced signal is not amplitude-modulated with the ISRF as the frequency of the modulation signal. The raw signal is simulated by Eq. (6.16) with f_r=20t+3 Hz, α=0, L_m=1, β=500, ω_r=4000*2π rad/s, ϕ_m=0, δ_m is a random number with a mean of 0.01, FCC=3.7, T=5s, N_1=3, f_I=4 f_r, B_1=[1.5, 1, 0.25], N_2=2, B_2=[0.5, 0.25], and SNR=-5 dB. The sampling rate is 20kHz and the sampling time is 5s. The simulated raw signal is shown in Fig. 6.7(a). The envelope TFR of the simulated "pure" bearing fault induced signal and the TFR of the simulated interference signal are shown in Fig. 6.7(b) and Fig. 6.7(c), respectively. From Fig. 6.7(b), the IFCF and its harmonics can be observed. From Fig. 6.7(c), the IIF, the ISRF, and their harmonics can be observed. Additionally, the envelope TFR of the raw signal is shown in Fig. 6.7(d), which is obtained by applying the Hilbert transform and the STFT to the raw signal. The IIF is dominant in the TFR in Fig. 6.7(d). Also, the ISRF can be observed even though it is not very clear. However, it is difficult to discern the IFCF from the IIF since

IFCF=3.7 f_r is so close to IIF which is 4 f_r. Obviously, it is impossible to detect the fault accurately directly from the raw signal.

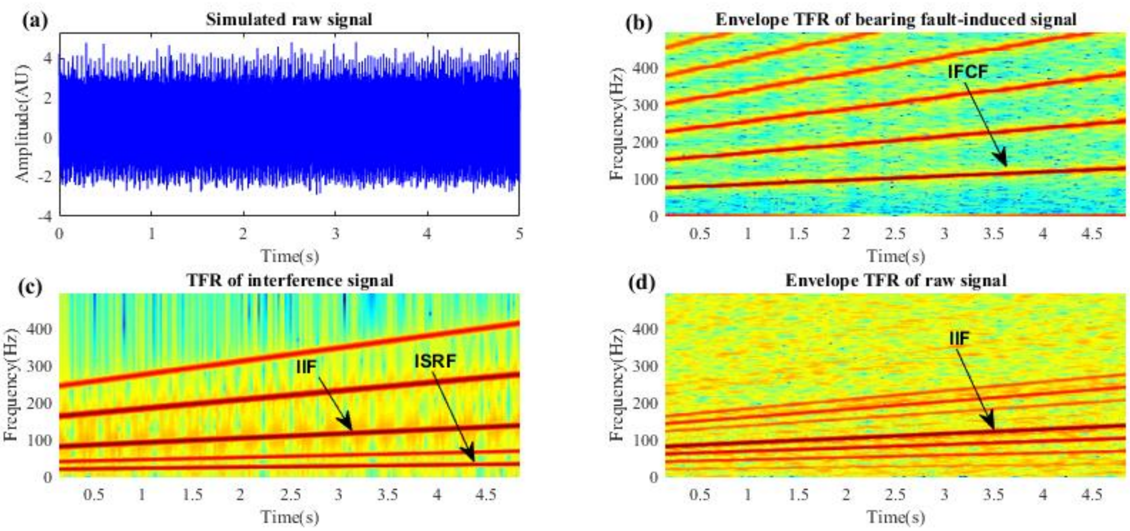

Fig. 6.7 Simulated signal for case 1. (a) Simulated raw signal, (b) envelope TFR of the bearing fault-induced signal, (c) TFR of the interference signal, and (d) envelope TFR of the raw signal.

The signal is then decomposed into a low oscillatory component, a high oscillatory component, and the residual via the OBSD, according to the proposed method. The parameters related to OBSD are chosen as Q_l =1, r_l =6, P_l =51, Q_h =9, r_h =6, P_h =230, $\lambda_l = \lambda_h$ =0.8, μ =2, and K =200 according to the guidance on parameter selection given in section 6.3.1. The low oscillatory component and the high oscillatory component are shown in Fig. 6.8(a) and Fig. 6.8(b), respectively. The decomposed low oscillatory component is taken as the extracted bearing fault signature. The envelope TRF of the low oscillatory component is shown in Fig. 6.8(c). It can be seen that only the IFCF and its harmonics appear in the TFR, and the IIF, the ISRF, and their harmonics disappear which implies the interference is completely removed. The TFR of the high oscillatory component is shown in Fig. 6.8(d). Only the IIF, the ISRF, and their harmonics show ridges in the TFR, the same as the TFR in Fig. 6.7(c), which also implies that the signal is effectively decomposed. Subsequently, the MTFCE algorithm is applied to the envelope TFR of the low oscillatory component and the 4 extracted curves are shown in Fig. 6.8(e). In addition, the average ratios are calculated as 1, 2, 3, and 4, respectively. None of the average ratio matches the given FCC (3.7) within a 5% relative error. The reason is that the

ISRF does not appear in the TFR shown in Fig. 6.8(c). Therefore, following the proposed ISRF-search fault diagnosis strategy, the MTFCE algorithm is applied to the TFR of the high oscillatory component and the 4 newly extracted curves are shown in Fig. 6.8(f). The extracted T-F curves are the ISRF and its 2nd harmonic, and the IIF and its 2nd harmonic. A new T-F curve figure is shown in Fig. 6.8(g), which is composed of four curves in Fig. 6.8(e) and the curve of the lowest frequency in Fig. 6.8(f). By taking the newly added curve in Fig. 6.8(g) as the bottom curve for the re-calculation of the average ratios, the average ratios for the 4 other curves in Fig. 6.8(f) are re-calculated as 3.70, 7.40, 11.11, and 14.80. Among these average ratios, 3.70 matches the given FCC (3.7) and the three other ratios match the 2nd, the 3rd, and the 4th harmonics, respectively. Thus, the bearing is diagnosed as faulty via the proposed method. Additionally, the new bottom curve in Fig. 6.8(g) is considered as the extracted ISRF and the curve with the second lowest frequency is considered as the extracted IFCF.

To compare the extracted ISRF to the preset ISRF and the extracted IFCF to the preset IFCF, a comparison figure is shown in Fig. 6.8(h). The average relative error of the extracted IFCF to the present IFCF is calculated as 0.36% and the average relative error of the extracted ISRF to the preset ISRF is calculated as 1.21%. It can be seen that both the extracted ISRF and the extracted IFCF fit the preset curves which further confirms the effectiveness of the proposed method.

To illustrate the superior performance of the OBSD method on bearing fault signature extraction under time-varying speed conditions, the widely used fast kurtogram is also applied to the simulated raw signal using the Matlab code provided by Antoni [112]. The fast kurtogram is a fast computational tool to determine the optimal band-pass filter by finding the filtered envelope with the maximum SK. The obtained kurtogram of the raw signal is shown in Fig. 6.9(a). From the kurtogram, the center frequency of the optimal filter is 469Hz and the bandwidth is 104Hz. The TFR of the envelope filtered by the optimal filter is shown in Fig. 6.9(b). Obviously, the IFCF cannot be observed in this TFR. The reason is that the optimal filter is not obtained as expected. For bearing signal filtering, the center frequency of the filter should be close to the high carrier frequency ω_r of the bearing fault-induced signal, which is 4000Hz in this simulation. The time-varying interferences contained in the signal may affect the performance of the SK method. Therefore, based on the results of this case study, the OBSD is superior to the band-pass filtering for bearing fault signature extraction under time-varying speed conditions.

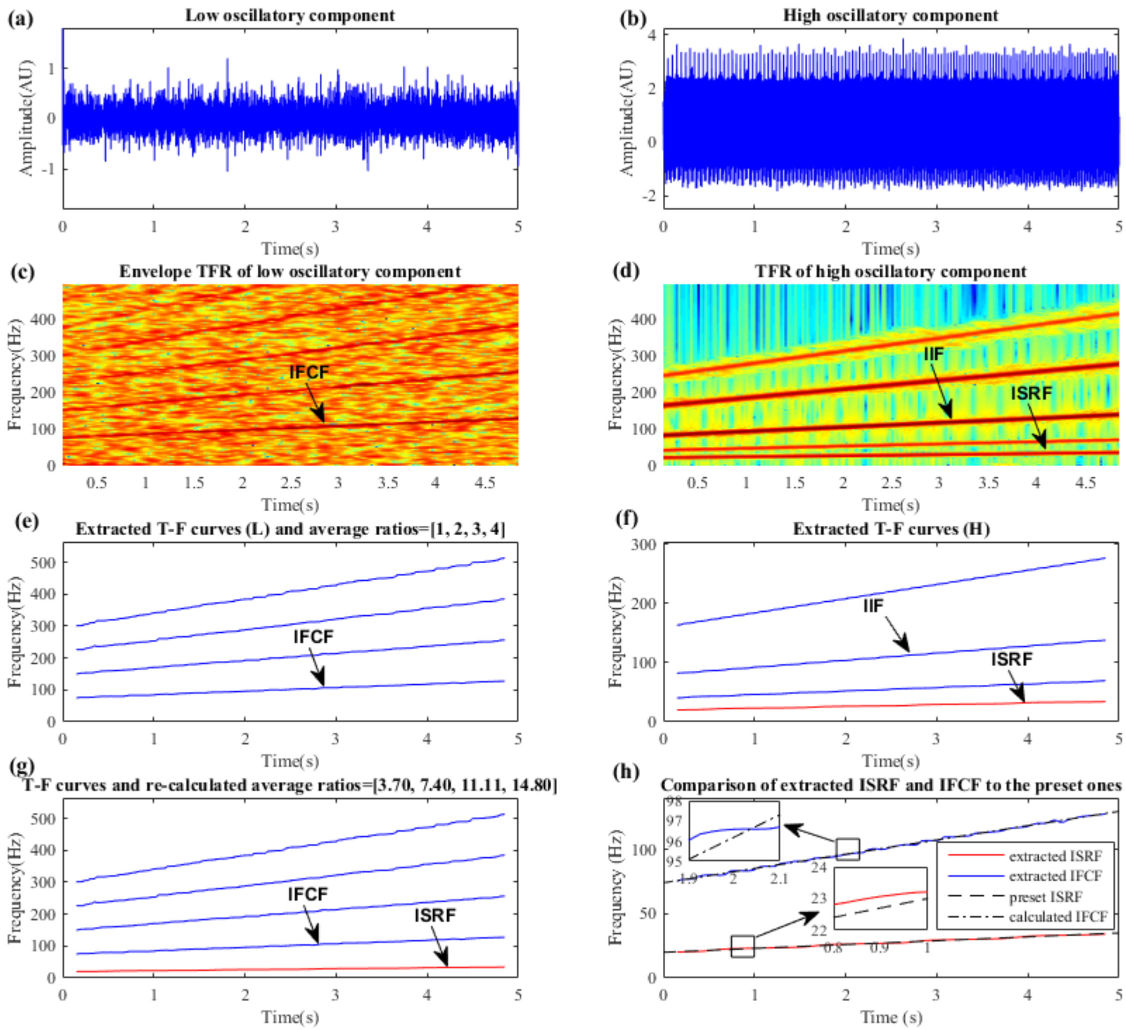

Fig. 6.8 Results of simulation case 1 via the proposed method. (a) Decomposed low oscillatory component via OBSD, (b) decomposed high oscillatory component, (c) envelope TFR of low oscillatory component, (d) TFR of high oscillatory component, (e) extracted T-F curves from (c); (f) extracted T-F curved from (d), (g) T-F curves used for bearing fault diagnosis, and (h) comparison of extracted ISRF and IFCF to their preset curves.

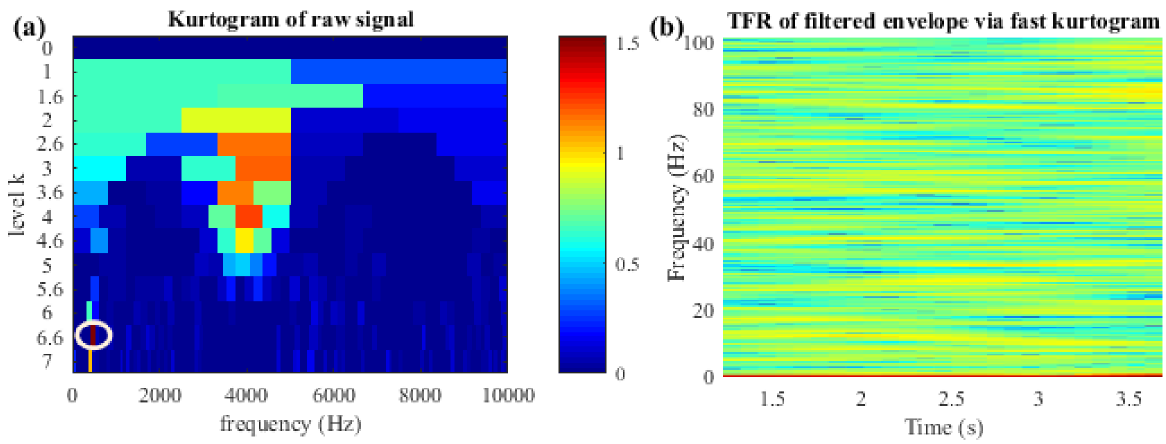

Fig. 6.9 Results of filtering via fast kurtogram for simulation case 1. (a) Kurtogram of the raw signal; (b) TFR of filtered envelope via fast kurtogram.

6.4.1.2 Case 2: ISRF is modulated by the bearing fault-induced signal

To further test the performance of the proposed fault diagnosis strategy, some changes were made to the simulated signal in case 1. In this simulation, the bearing fault-induced signal is amplitude modulated with the ISRF as the frequency of the modulation signal, i.e., the ISRF will appear in the envelope TFR of bearing fault-induced signal, which is similar to the case of bearing signal with inner race fault. The original signal is simulated by Eq. (6.16) with f_r =20t+3 Hz, α =0.95, L_m =1, β =500, ω_r =4000*2π rad/s, ϕ_m =0, δ_m is random number with a mean of 0.01, FCC=3.7, T=5s, f_I = 4 f_r, N_1 =3, B =[1.5, 1, 0.25], N_2 =2, B_2 =[0.5, 0.25], and SNR=-5 dB. The sampling rate is 20kHz and the sampling time is 5s. The only difference from the simulation in case 1 is that α is changed from 0 to 0.95. The simulated raw signal is shown in Fig. 6.10(a), which is a mixture of bearing fault-induced signal, two set of interferences and noise. In addition, the envelope TFR of the raw signal is shown in Fig. 6.10(b), in which the IIF can be clearly observed. However, the IFCF cannot be observed in the envelope TFR which implies that the bearing fault cannot be detected and diagnosed without further analysis.

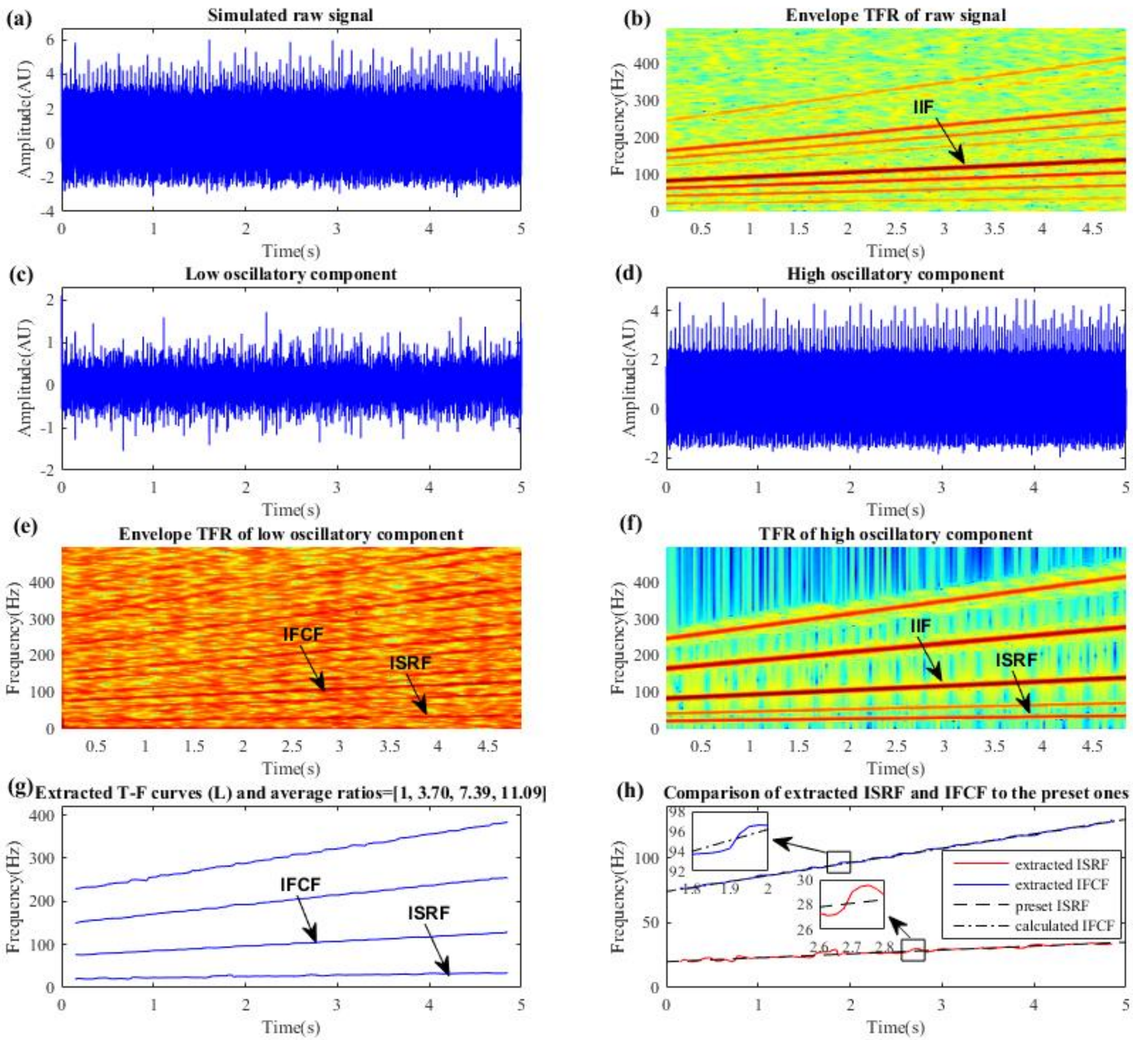

Fig. 6.10 Simulated signal and results for case 2. (a) Simulated raw signal, (b) envelope TFR of raw signal, (c) decomposed low oscillatory component via OBSD, (d) decomposed high oscillatory component, (e) envelope TFR of low oscillatory component, (f) TFR of high oscillatory component, (g) extracted T-F curves from (e), and (h) comparison of extracted ISRF and IFCF to their preset curves

Applying the proposed method to the simulated raw signal, the decomposed low oscillatory component shown in Fig. 6.10(c), the high oscillatory component shown in Fig. 6.10(d) are obtained via the OBSD. Parameters related to the OBSD are selected as Q_l =1, r_l =6, P_l =51, Q_h =7, r_h =6, P_h =189, $\lambda_l = \lambda_h$ =0.7, μ =2, and K =250 according to the discussions in section 6.3.1. The envelope TFR of the low oscillatory component is shown in Fig. 6.10(e) and the TFR of the high oscillatory component is shown in Fig. 6.10(f). The IFCF and its harmonics, the ISRF and sidebands can be observed in Fig. 6.10(e) without the presence of IIF and its harmonics, which implies that the bearing fault signature is effectively extracted from the contaminated raw signal via the OBSD. Additionally, the IIF and the ISRF, and their harmonics are clearly shown in Fig. 6.10(f), which also demonstrates that the signal is effectively separated. By applying the MTFCE algorithm to the envelope TFR shown in Fig. 6.10(e), four T-F curves are extracted as shown in Fig. 6.10(g). Taking the curve of the lowest frequency as the bottom curve, the average ratios for the 4 curves are calculated as 1, 3.70, 7.39, and 11.09, respectively. Among these average ratios, 3.70 matches exactly the given FCC (3.7). Additionally, 7.39 and 11.09 imply the 2nd and the 3rd harmonics of IFCF. Based on the result, the bearing fault can be diagnosed. Also, the bottom curve in Fig. 6.10(g) is considered as the extracted ISRF and the curve with the second lowest frequency is considered as the extracted IFCF.

A comparison figure which includes the extracted ISRF, the present ISRF, the extracted IFCF, and the preset IFCF is shown in Fig. 6.10(h). It can be observed that both the extracted ISRF and the extracted IFCF fit the preset curves very well. The average relative error of the extracted IFCF to the preset IFCF is 0.46% and the average relative error of the extracted ISRF to the preset ISRF is 1.80%, which demonstrates the effectiveness of the proposed method.

6.4.2 Cases when IFCF and IIF cross each other

To test the performance of the proposed method for different situations, two additional simulations are conducted in which the IFCF of the bearing fault-induced signal and the IIF of the interference cross each other. Theoretically, the OBSD should be able to separate the bearing signal from the interference in this situation since it is independent of frequency. This will be demonstrated in the following simulations.

6.4.2.1 Case 3: linearly varying IFCF and IIF

In this simulation, the IFCF and IIF cross each other and both are linearly varying. Additionally, the ISRF is not modulated by the bearing fault-induced signal. Similarly, the interference signal is composed of two sets of interferences and one of them with ISRF as the fundamental frequency. The raw signal is simulated by Eq. (6.15) with f_r=20t+3 Hz, α=0, L_m=1, β=500, ω_r=4000*2π rad/s, ϕ_m=0, δ_m is random number with a mean of 0.01, FCC=3.7, T=10s, N_1=3, f_I= -7t+210, B_1=[1.5, 1, 0.25], N_2=2, B_2=[0.5, 0.25] and SNR=-5 dB. The sampling rate is 20kHz and the sampling time is 5s. The simulated signal raw signal is shown in Fig. 6.11(a). The envelope TFR of the simulated bearing fault-induced signal is shown in Fig. 6.11(b), in which the IFCF and its harmonics can be observed. The TFR of the interference signal is shown in Fig. 6.11(c) which is dominated by the ridges of the interferences, i.e. the IIF and its two harmonics, and the ISRF and its 2nd harmonic. Moreover, the envelope TFR of the raw signal is shown in Fig. 6.11(d), in which the IFCF can barely be observed and it crosses the IIF. Again, the bearing fault cannot be detected directly without further investigation.

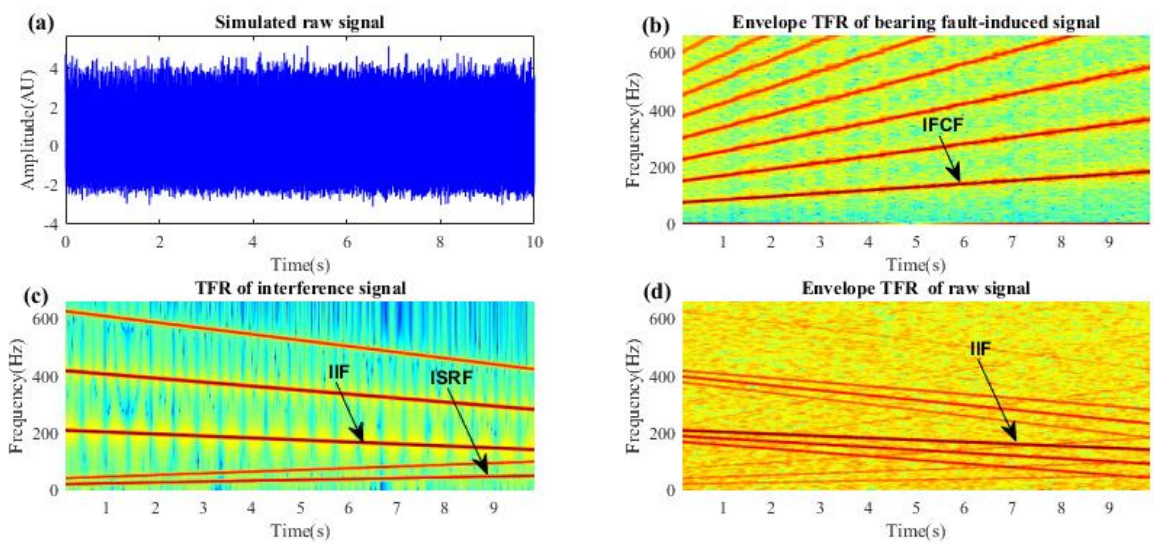

Fig. 6.11 Simulated signal for case 3. (a) Simulated raw signal, (b) envelope TFR of the bearing fault-induced signal, (c) TFR of the interference signal, and (d) envelope TFR of the raw signal.

The OBSD is then applied to the raw signal to separate the bearing fault signature and interferences with parameters selected as Q_l=1, r_l=6, P_l=51, Q_h=8, r_h=6, P_h=210, $\lambda_l=\lambda_h$=0.5, μ=1, and K=200, according to the guidance given in section 6.3.1. The decomposed low oscillatory component is shown in Fig. 6.12(a) and the high oscillatory component is shown in Fig. 6.12(b), respectively. The envelope TFR of the low oscillatory component is shown in Fig.

6.12(c) in which the IFCF and its harmonics can be clearly observed without the presence of the T-F curves of interferences. The TFR of the high oscillatory component is shown in Fig. 6.12(d) in which only the T-F curves of interferences can be seen, i.e. the IIF and its two harmonics, the ISRF and its 2nd harmonic. The results shown in Fig. 6.12(c) and Fig. 6.12(d) reveal that the bearing fault signature and the interference signal are effectively separated by the OBSD. According to the proposed method, the MTFCE algorithm is applied to the envelope TFR of the low oscillatory component and the extracted 4 curves are presented in Fig. 6.12(e). The average ratios are calculated as 1, 2, 3, and 4, respectively. Similar to the results in case 2, the extracted T-F curves are the IFCF and its harmonics. The bearing fault still cannot be diagnosed since none of the calculated average ratios match the given FCC (3.7). Therefore, according to the proposed ISRF-search diagnosis strategy, 4 extra T-F curves are extracted from the TFR of the high oscillatory component shown in Fig. 6.12(f) via the MTFCE algorithm. A new T-F figure is obtained as shown in Fig. 6.12(g) by adding the curve which has the lowest frequency in Fig. 6.12(f) to Fig. 6.12(e). The average ratios are re-calculated as 3.7, 7.4, 11.1, and 14.8, respectively. Among these average ratios, 3.7 matches the FCC and the rest of the ratios match the 2nd, 3rd, and 4th harmonic, respectively. Now the bearing fault can be successfully diagnosed. Additionally, the bottom curve and the curve with the second lowest frequency in Fig. 6.12(g) are considered as the extracted ISRF and the extracted IFCF, respectively.

To further examine the performance of the proposed method, the extracted ISRF, extracted IFCF, and their preset curves are plotted in Fig. 6.12(h). It can be seen that both the extracted ISRF and the extracted IFCF tightly fit the preset curves. The average relative error of the extracted IFCF to the preset IFCF is only 0.36% and the average relative error of the extracted ISRF to the preset ISRF is 0.93%. The comparison further testifies to the effectiveness of the proposed method.

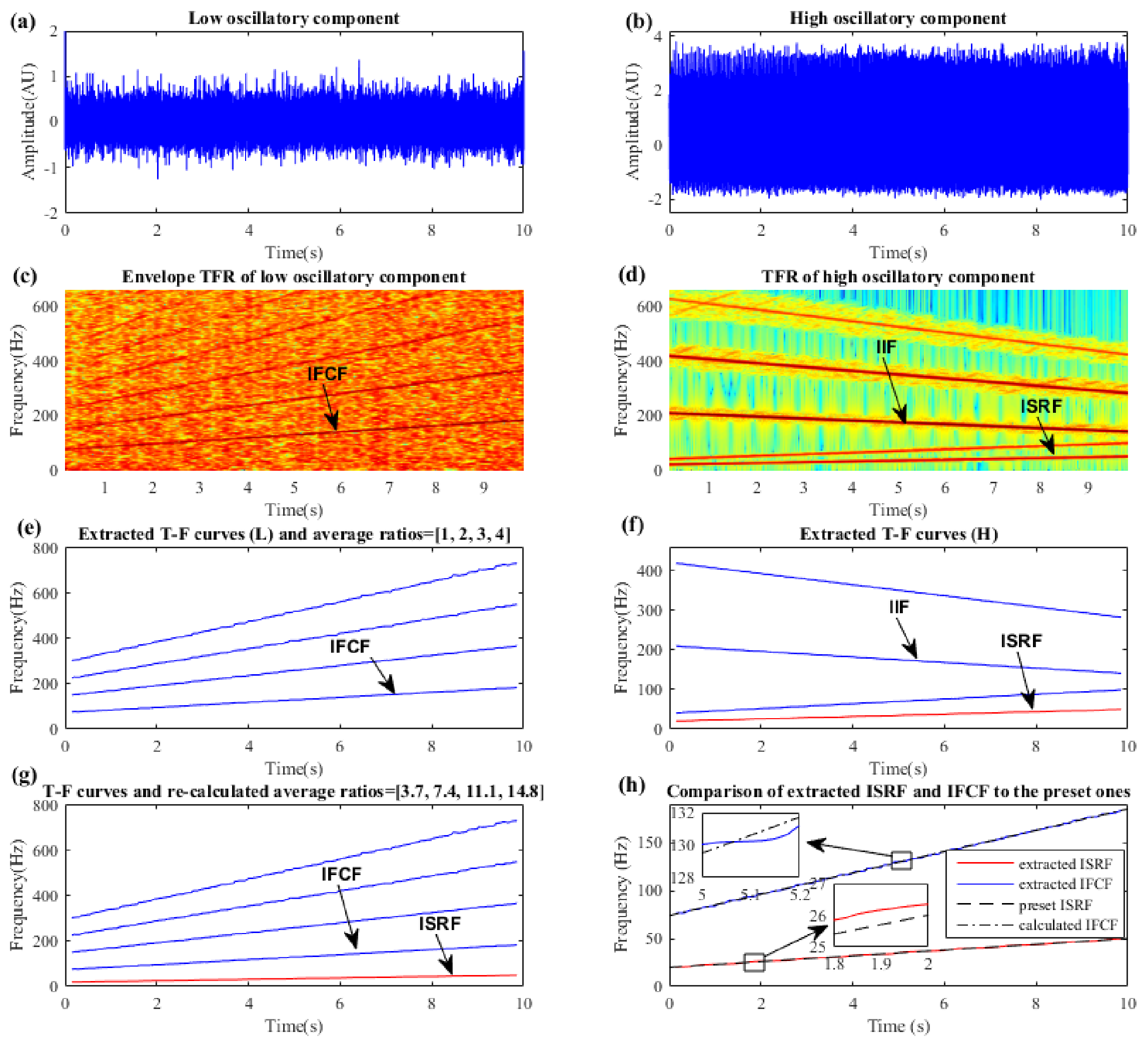

Fig. 6.12 Results of simulation case 3. (a) Decomposed low oscillatory component via OBSD, (b) decomposed high oscillatory component, (c) envelope TFR of low oscillatory component, (d) TFR of high oscillatory component, (e) extracted T-F curves from (c), (f) extracted T-F curved from (d); (g) T-F curves used for bearing fault diagnosis, and (h) comparison of extracted ISRF and IFCF to their preset curves.

6.4.2.2 Case 4: linearly varying IFCF and non-linearly varying IIF

To further challenge the proposed method, non-linearly varying IIF of the interference is considered in this simulation while keeping linearly varying IFCF of the bearing fault-induced signal. Again, the IFCF and the IIF cross each other. Additionally, the bearing fault-induced signal is amplitude modulated with the ISRF as the frequency of the modulation signal. The raw signal is simulated by Eq. (6.15) with f_r =20t+3 Hz, α =0.95, L_m =1, β =500, ω_r =4000*2π rad/s, ϕ_m =0, δ_m is a random number with a mean of 0.01, FCC=3.7, T=10s, N_1 =3, f_I = -6(t-5)^2+250, B_1 =[1.5, 1, 0.25], N_1 =2, B_2 =[0.5, 0.25], and SNR=-5 dB. The sampling rate is 20kHz and the sampling time is 5s. The simulated raw signal is shown in Fig. 6.13(a). The envelope TFR of the raw signal is shown in Fig. 6.13(b) in which the IFCF can only be vaguely observed, and the IFCF and the IIF cross each other. From the TFR in Fig. 6.13(b), the bearing fault once again cannot be diagnosed directly.

Therefore, the OBSD is applied to the raw signal. The low oscillatory component is shown in Fig. 6.13(c) and the high oscillatory component is shown in Fig. 6.13(d). Parameters related to the OBSD method are selected as Q_l =1, r_l =6, P_l =51, Q_h =9, r_h =6, P_h =251, $\lambda_l = \lambda_h$ =0.7, μ =1, and K =200 according to the guiding principles in section 6.3.1. The envelope TFR of the low oscillatory component is shown in Fig. 6.13(e). The ISRF, the IFCF, and its harmonics can be observed in Fig. 6.13(e) without the appearance of the IIF. Also, the TFR of the high oscillatory component is shown in Fig. 6.13(f) in which only the IIF, the ISRF, and their harmonics appear. It indicates that the OBSD has effectively separated the bearing fault signature and the interference. Subsequently, the MTFCE algorithm is applied to the TFR shown in Fig. 6.13(e) and the extracted curves are shown in Fig. 6.13(g). The calculated average ratios are 1, 3.71, 7.42, and 11.13, respectively. Among these average ratios, 3.71 matches the FCC with very low relative error (0.27%). Additionally, 7.42 and 11.13 correspond to the 2nd harmonic and the 3rd harmonic, respectively. Thus, the bearing fault can be diagnosed via the proposed method. Furthermore, a comparison figure which includes the extracted ISRF, the preset ISRF, the extracted IFCF, and the preset IFCF is shown in Fig. 6.13(h). It can be seen that both the extracted IFCF and the extracted ISRF fit their corresponding preset curves well. In addition, the average relative error of the extracted IFCF to the preset IFCF is calculated as 0.34% and the average relative error of the extracted ISRF to the preset ISRF is 2.22%. The comparison also confirms the effectiveness of the proposed method in handling such a complicated case.

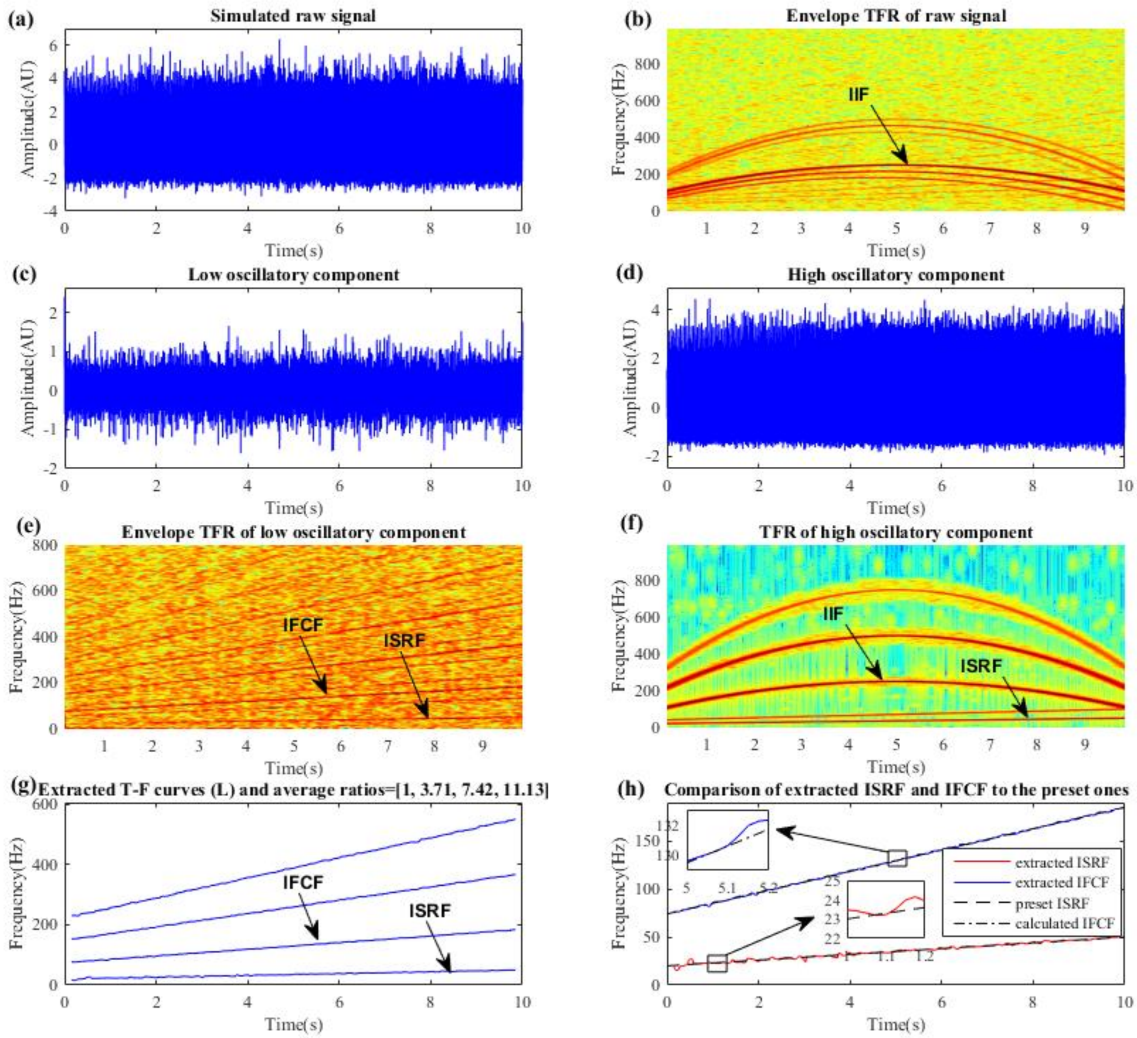

Fig. 6.13 Results of simulation case 4. (a) Simulated raw signal, (b) envelope TFR of raw signal, (c) decomposed low oscillatory component via OBSD, (d) decomposed high oscillatory component, (e) envelope TFR of low oscillatory component, (f) TFR of high oscillatory component, (g) extracted T-F curves from (e), and (h) comparison of extracted ISRF and IFCF to their preset curves

6.5 Experimental validation

To further test the performance of the proposed method, it is applied to signals collected from experiments. Experiments are conducted on a SpectraQuest machinery fault simulator (MFS-PK5MT) to collect the bearing vibration signal which is contaminated by interference transmitted from a gearbox and noise. To validate the effectiveness of the proposed method, two different types of signals are collected. In one experiment, the shaft and gearbox are driven by the same motor, i.e. the IFCF is proportional to the gear meshing frequency (fundamental frequency of the interference). In the other experiment, the shaft and gearbox are driven by two different motors, and hence the IFCF and the gear meshing frequency are independent.

6.5.1 Experiment 1: Shaft and gearbox driven by the same motor

The set-up of this experiment is shown in Fig. 6.14. The shaft is supported by two bearings and one of them is a faulty bearing with an outer race fault. The shaft is driven by a motor and the motor is controlled by an AC drive. A gearbox is connected to the shaft by a belt. Parameters of bearings and gears used in this experiment are given in Table 6.1. The FCC for an outer race fault (FCC_O) is calculated as 3.57 by Eq. (6.9) and the FCC for inner race fault (FCC_I) is calculated as 5.43 by Eq. (6.10). The gear meshing frequency is (18/2.6)=6.92 times the shaft rotational frequency. A sensor (accelerometer) is mounted on the base of the test rig to collect the vibration signal. Therefore, the collected signal contains not only the bearing vibration signal but also the gear meshing signal. It should be pointed out that to make the fault detection more challenging, the accelerometer is purposely mounted close to the gearbox yet far away from the shaft that is supported by the faulty bearing. In this way, the interference effect is enhanced relative to the faulty bearing signal. The signal is sampled by Labview with sampling frequency 20kHz and the sampling time is 4.46s. Additionally, to verify the results obtained by the proposed method, a tachometer is used to measure the time-varying shaft rotational speed.

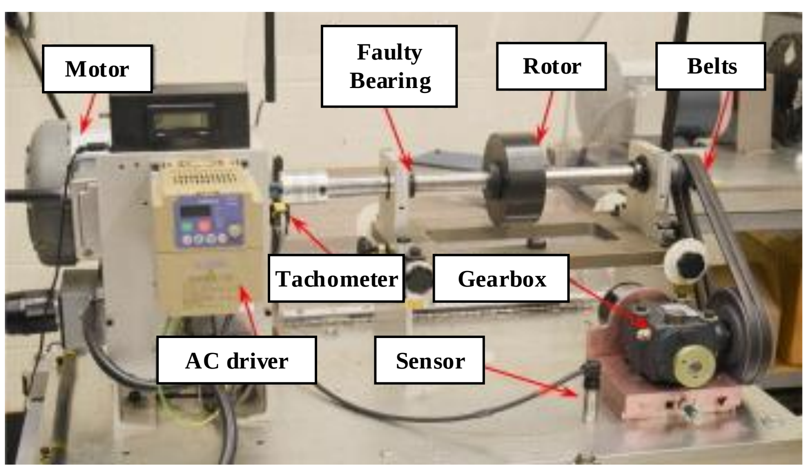

Fig. 6.14 Set-up for Experiment 1

Table 6.1 Parameters of bearing and gearbox

Bearing type	Number of balls	Pitch diameter	Ball diameter
ER16K	9	38.52mm	7.94mm
Bearing FCC_O	Bearing FCC_I	Diameter ratio of sheaves	Number of gear teeth
3.57	5.43	1:2.6	18

The collected raw signal is shown in Fig. 6.15(a) and the measured ISRF is shown in Fig. 6.15(b) which increases from 30.25Hz to 60.5Hz. The TFR of the raw signal is obtained via the STFT, as shown in Fig. 6.15(c). It can be seen that the TFR of the raw signal is dominated by the instantaneous gear meshing frequency, i.e. the IIF. The ISRF can also be seen in Fig. 6.15(c), but the IFCF and its harmonics cannot be observed. Moreover, in the envelope TFR of the raw signal, shown in Fig. 6.15(d), no clear T-F curves can be observed. Obviously, the bearing fault cannot be detected and diagnosed with the raw signal directly.

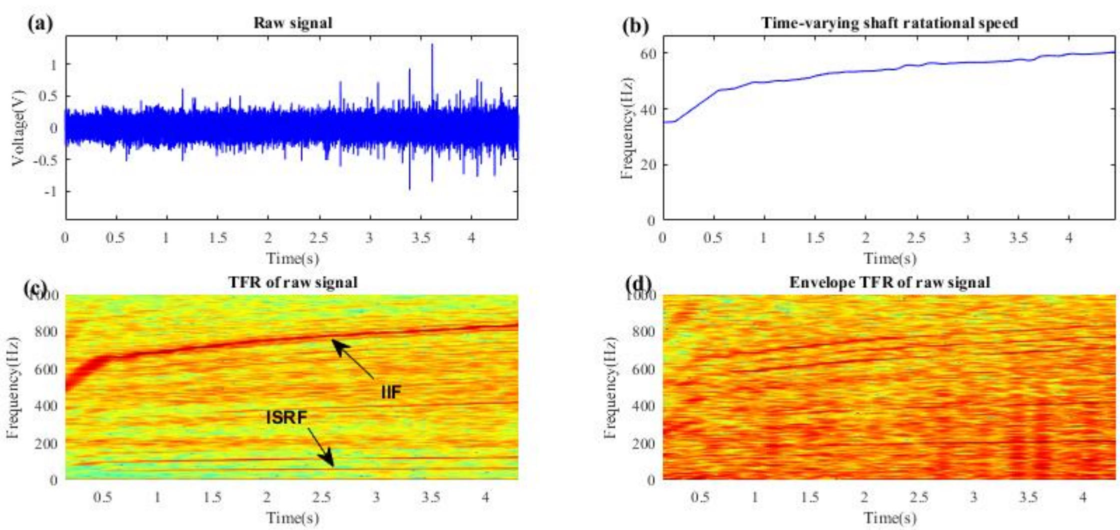

Fig. 6.15 Collected signal of Experiment 1. (a) Collected raw signal; (b) measured ISRF; (c) TFR of the raw signal; (d) envelope TFR of the raw signal.

According to the proposed method, the OBSD is applied to the raw signal with parameters Q_l =1, r_l =6, P_l =51, Q_h =8, r_h =6, P_h =210, $\lambda_l = \lambda_h$ =0.1, μ =1, and K =250 according to the guidance on the parameter selection given in section 6.3.1. The decomposed low oscillatory component is shown in Fig. 6.16(a) and the high oscillatory component is shown in Fig. 6.16(b), respectively. The envelope TFR of the low oscillatory component is obtained via the Hilbert transform and STFT, shown in Fig. 6.16(c). The IFCF and its harmonics can be observed in Fig. 6.16(c) without the presence of the IIF and its harmonics. The TFR of the high oscillatory component is also obtained (Fig. 6.16 (d)) in which the IIF and the ISRF can be observed without the IFCF. The results in Fig. 6.16(c) and Fig. 6.16(d) demonstrate that the OBSD has effectively separated the bearing fault signature and the interference. Next, the MTFCE algorithm is applied to the envelope TFR of the low oscillatory component, four curves are extracted as shown in Fig. 6.16(e), and the average ratios are calculated as 1, 1.67, 3.76, and 7.58. However, none of the average ratio match the FCC_O (3.57) or the FCC_I (5.43). Hence, according to the proposed method, four additional T-F curves are extracted from the TFR of the high oscillatory component shown in Fig. 6.16(f) via the MTFCE algorithm. Adding the curve of the lowest frequency in Fig. 6.16(f) to Fig. 6.16(e), a new figure of T-F curves is generated shown in Fig. 6.16(g). Average ratios are re-calculated as 0.94, 1.57, 3.49 and 7.06 by taking the newly added curve as the bottom curve. Among these new average ratios, 3.49 matches the FCC_O (3.57) with relative error 2.24% which is within 5%. Thus, the curve of the second highest frequency in Fig. 6.16(e) is considered as the extracted IFCF and the curve of the lowest curve in Fig. 6.16(f) is considered as the extracted ISRF. Furthermore, 7.06 matches 2* FCC_O which is associated with the 2nd harmonic of the IFCF. Therefore, an outer race fault is diagnosed for the faulty bearing using the proposed method.

For comparison, the extracted ISRF, the measured ISRF, the extracted IFCF and the calculated IFCF are shown in Fig. 6.16(h). It can be observed that both the ISRF and the extracted IFCF fit the measured curve or the calculated curve quite well. The average relative error of the extracted ISRF to the measured ISRF and the average relative error of the extracted IFCF to the measured IFCF are calculated as 2.54% and 2.32%, respectively.

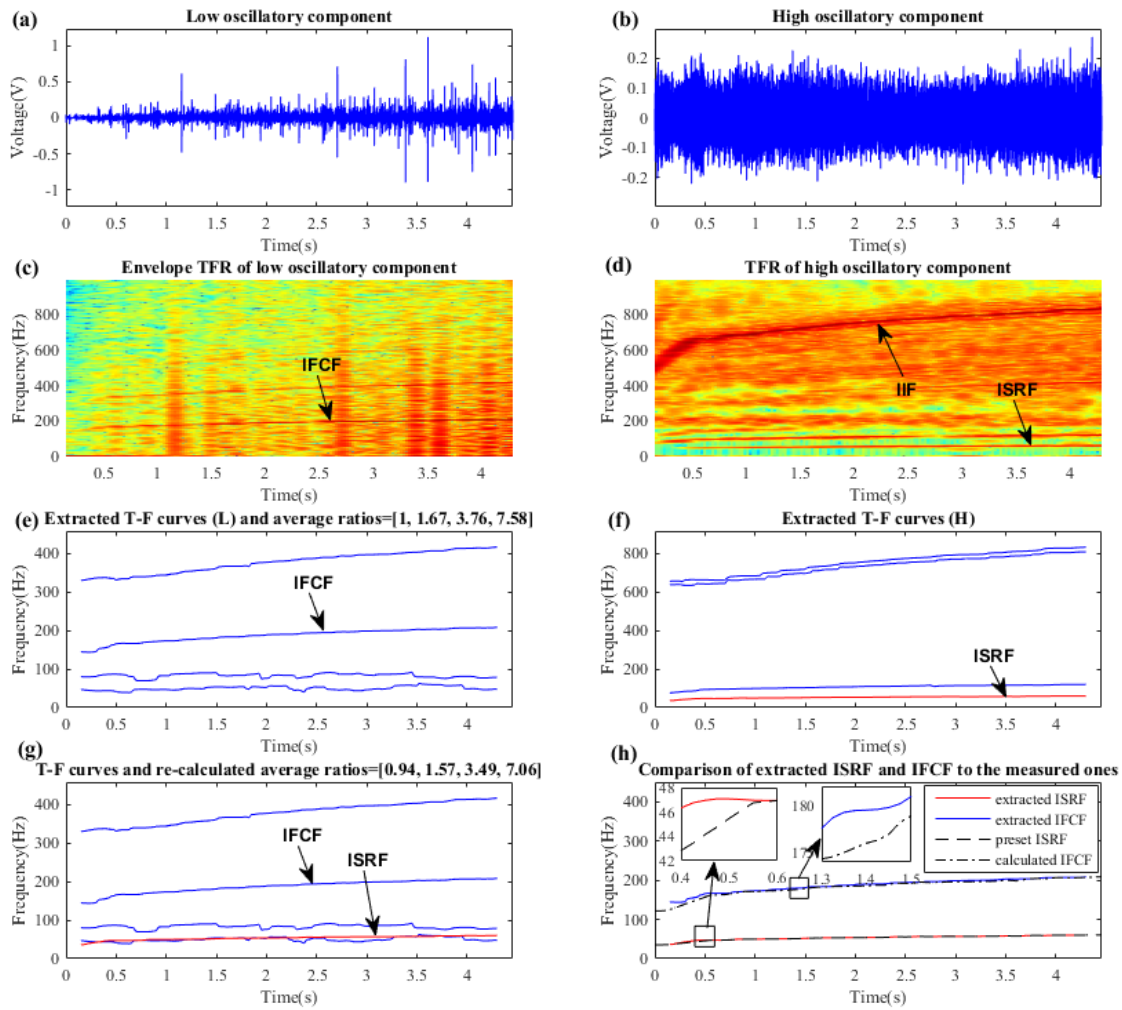

Fig. 6.16 Results of Experiment 1. (a) Decomposed low oscillatory component via OBSD, (b) decomposed high oscillatory component, (c) envelope TFR of low oscillatory component, (d) TFR of high oscillatory component, (e) extracted T-F curves from (c), (f) extracted T-F curved from (d), (g) T-F curves used for bearing fault diagnosis, and (h) comparison of extracted ISRF and IFCF to their preset curves.

To compare the performance of the OBSD method to band-pass filtering method on bearing fault signature extraction under time-varying speed conditions, the fast kurtogram is applied to the collected signal. The kurtogram of the raw signal is shown in Fig. 6.17(a), obtained using the Matlab code provided by Antoni [112]. The center frequency of the optimal filter is obtained as 6250Hz and the bandwidth is 833Hz. With the optimal filter, the TFR of the filtered envelope is shown in Fig. 6.17(b). The IFCF can be observed in the TFR, however, its 2nd harmonic can barely be observed. The interferences are removed since the IIF and the ISRF disappear in the TFR. However, compared to Fig. 6.16(c), the result obtained from the proposed OBSD method is much more convincing since the 2nd harmonic of the IFCF can be more clearly observed in the TFR of the low oscillatory component obtained via the OBSD. The T-F curves extracted from the TFR of the filtered envelope via the MTFCE are shown in Fig. 6.17(c). Among the four curves, the IFCF is extracted with an average relative error of 3.11% to the measured IFCF, which is higher than the average relative error of 2.32% obtained in the comparison shown in Fig. 6.16(h). Additionally, the 2nd harmonic of the IFCF cannot be extracted from the TFR because it is not clear in Fig. 6.17(b). Therefore, the performance of the OBSD is better than the fast kurtogram for bearing fault signature extraction under time-varying speed conditions. Moreover, the interference signal can also be obtained via the OBSD which is also useful for the extraction of ISRF from the TFR.

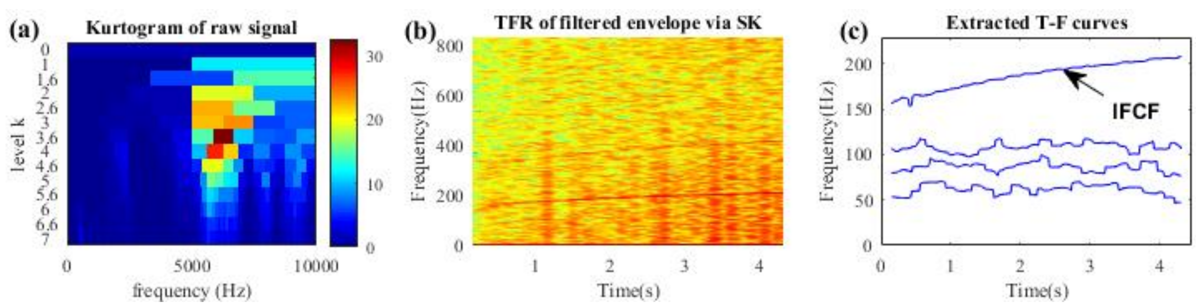

Fig. 6.17 Results of filtering via fast kurtogram for experiment 1. (a) Kurtogram of the raw signal, (b) TFR of filtered envelope via fast kurtogram, and (c) extracted T-F curves from (b).

6.5.2 Experiment 2: Shaft and gearbox driven by different motors

In this case, the set-up of the experiment is shown in Fig. 6.18. The only difference from the first experiment (Fig. 6.14) is that the gearbox is connected to another shaft driven by another motor, labeled as motor 2 in Fig. 6.18. Thus, the gear meshing frequency is 6.92 times the rotational frequency of the shaft that the gearbox connects to. The FCC_O (3.57) and the FCC_I (5.43) are the same since the bearings are the same. Once again, the accelerometer is intentionally placed in the vicinity of the gearbox but far away from the faulty bearing so that the interference signal will be more pronounced. The sampling frequency is 20kHz and the sampling time is 6.23s.

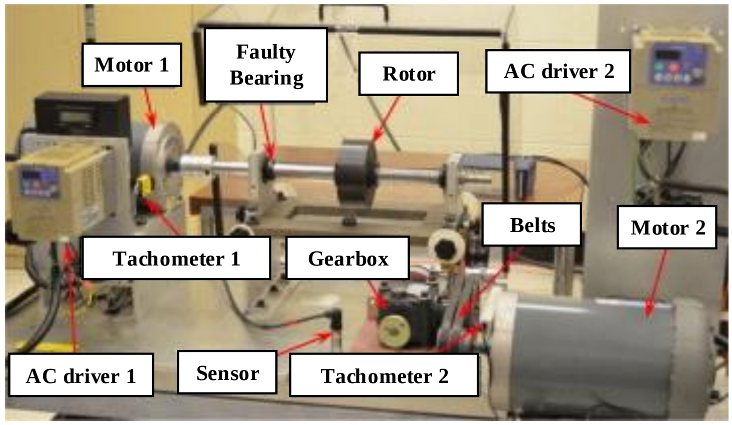

Fig. 6.18 Set-up of Experiment 2

The collected raw signal is shown in Fig. 6.19(a) and the measured ISRF of the shaft supported by the faulty bearing is shown in Fig. 6.19(b) which rises almost linearly from 23Hz to 36Hz. The TFR of the raw signal is shown in Fig. 6.19(c). Similarly, the TFR is dominated by the IIF (gear meshing frequency) and the ISRF also appears. The envelope TFR of the raw signal is shown in Fig. 6.19(d), in which only the IIF and its harmonics can be observed. It can be seen that in both Fig. 6.19(c) and Fig. 6.19(d), the IFCF and its harmonics are absent which again indicates that the bearing fault cannot be directly detected from the raw signal.

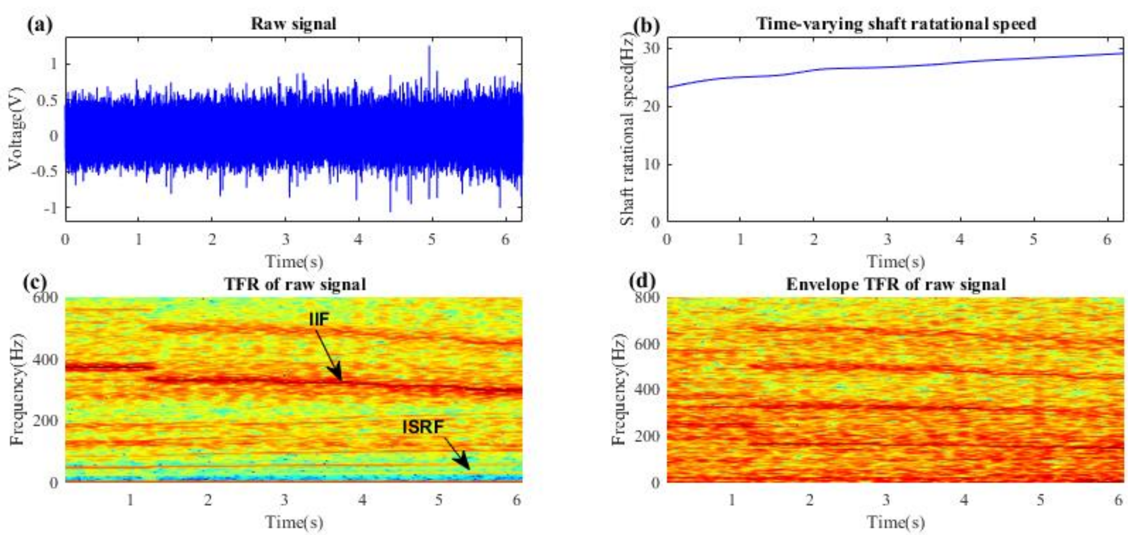

Fig. 6.19 Collected signal of Experiment 2. (a) Collected raw signal; (b) measured ISRF; (c) TFR of the raw signal; (d) envelope TFR of the raw signal.

To extract the bearing fault signature from the obscured raw signal, the OBSD is applied to the raw signal with parameters Q_l =1, r_l =6, P_l =51, Q_h =9, r_h =6, P_h =251, $\lambda_l = \lambda_h$ =0.1, μ =2, and K =200 according to the discussions in section 6.3.1. The obtained low oscillatory component and the high oscillatory component are shown in Fig. 6.20(a) and Fig. 6.20(b), respectively. The envelope TFR of the low oscillatory component is displayed in Fig. 6.20(c). The IFCF and its 2nd harmonic can be clearly observed without the presence of the IIF, which implies that the bearing fault signature is effectively extracted as the low oscillatory component. Additionally, the TFR of the high oscillatory component is shown in Fig. 6.20(d) where only the IIF and the ISRF can be observed, without the presence of the IFCF, which indicates that the OBSD has successfully separated the bearing fault signature and the interference. According to the proposed method, four T-F curves are extracted from the TFR, shown in Fig. 6.20(c), via the MTFCE algorithm. The extracted T-F curves are shown in Fig. 6.20(e). The calculated average ratios for the four curves are 1, 1.90, 4.10 and 8.24, respectively. None of them match the FCC_O (3.57) or the FCC_I (5.43) since the ISRF is not present in Fig. 6.20(c). Thus, another four curves are extracted (shown in Fig. 6.20(f)) from the TFR shown in Fig. 6.20(d), in accordance with the proposed method. A new figure is generated as shown in Fig. 6.20(g) by adding the curve with the lowest frequency in Fig. 6.20(f) to the T-F curves in Fig. 6.20(f). By taking the newly added curve as the bottom curve, the average ratios of the four other curves in Fig. 6.20(g) are calculated as 0.90, 1.63, 3.54 and 7.07. Among these average ratios, 3.54 matches the FCC_O (3.57) with relative error 0.84% (within 5%). Therefore, the new bottom curve and the second highest curve in Fig. 6.20(g) are treated as the extracted ISRF and the extracted IFCF, respectively. In addition, 7.07 matches 2* FCC_O which indicates that the highest curve in Fig. 6.20(e) is the 2nd harmonic of the IFCF. Hence, the bearing can be diagnosed as faulty with an outer race fault via the proposed method.

To further demonstrate the performance of the proposed method, a comparison figure of the extracted ISRF to the measured ISRF and the extracted IFCF to the calculated IFCF is shown in Fig. 6.20(h). Similarly, it can be seen that both the extracted ISRF and the extracted IFCF fit the measured curve or the calculated curve. Also, the average relative error of the extracted ISRF to the measured ISRF and the average relative error of the extracted IFCF to the calculated IFCF are calculated as 2.04% and 1.53%, respectively, which confirms the effectiveness of the proposed method.

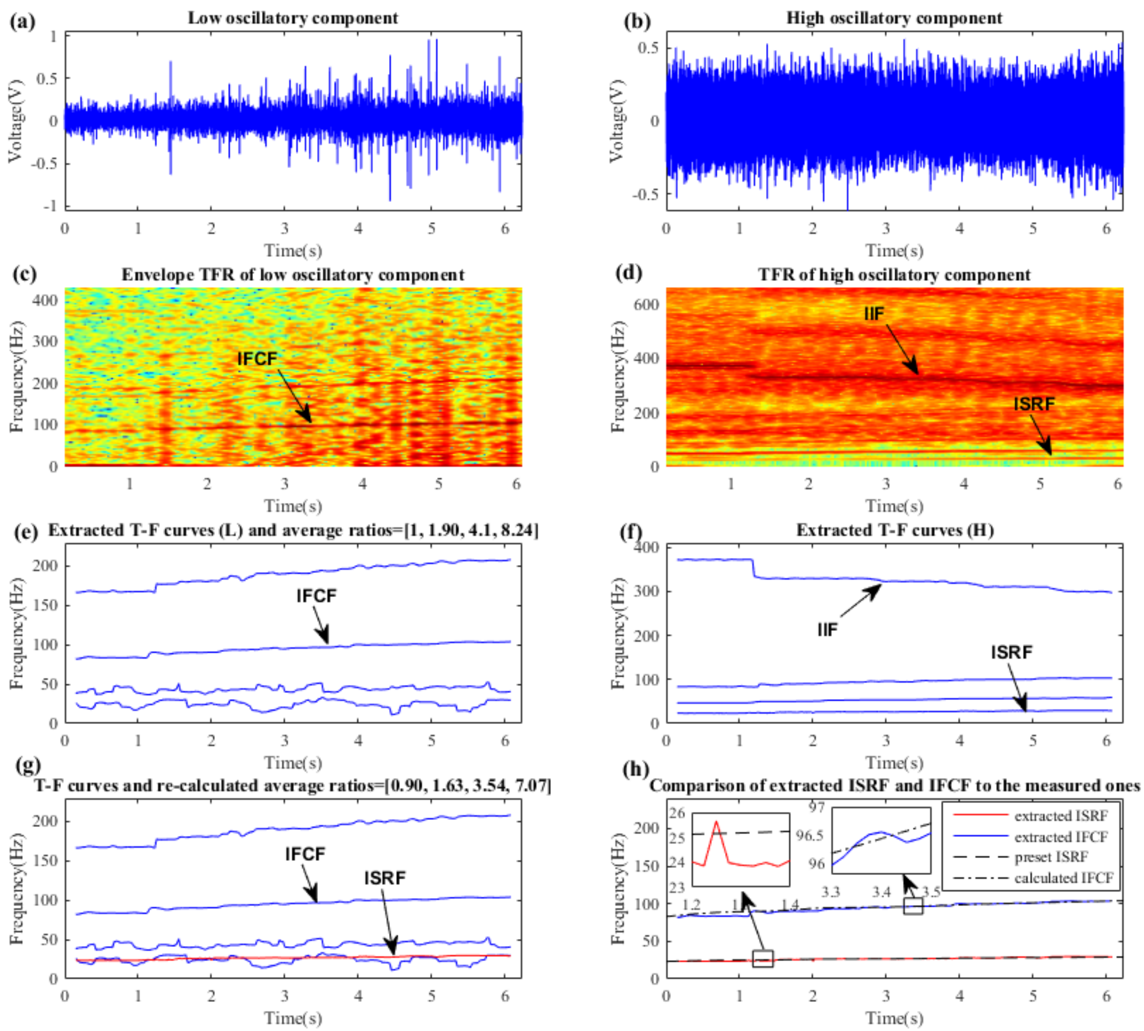

Fig. 6.20 Results of Experiment 2. (a) Decomposed low oscillatory component via OBSD, (b) decomposed high oscillatory component, (c) envelope TFR of low oscillatory component, (d) TFR of high oscillatory component, (e) extracted T-F curves from (c), (f) extracted T-F curved from (d), (g) T-F curves used for bearing fault diagnosis, and (h) comparison of extracted ISRF and IFCF to their preset curves.

Similarly, the fast kurtogram is applied to the raw signal to compare to the results obtained via the OBSD. The kurtogram of the raw signal is shown in Fig. 6.21(a), implemented via Matlab code provided by Antoni [112]. The optimal band-pass filter is obtained as center frequency=7708Hz and bandwidth=417Hz. Then, the TFR of the filtered envelope is shown in Fig. 6.21(b). Similar to the TFR shown in Fig. 6.17(b), the IFCF is visible but its 2nd harmonic is barely visible. Both the IIF and the ISRF are absent from the TFR, which implies the interferences are canceled. However, the ISRF is important for bearing fault diagnosis (e.g., to identify if the fault originates from an inner or outer race). The four extracted T-F curves obtained via the MTFEC are shown in Fig. 6.21(c). Again, the IFCF is extracted with an average relative error of 1.79% to the calculated IFCF, which is higher than the average relative error (1.53%) of the IFCF in the comparison shown in Fig. 6.20(h). The 2nd harmonic of the IFCF is not extracted due to its poor "readability" in Fig. 6.21(b). However, the 2nd harmonic of the IFCF is successfully extracted from the envelope TFR of the low oscillatory component obtained via the OBSD. Also, compared to the results obtained via the OBSD, the OBSD can extract not only the bearing fault signature but also the interference signal. The ISRF can still be extracted from the TFR of the obtained high oscillatory component even if the ISRF does not appear in the envelope TFR of the extracted bearing fault signature (i.e. the low oscillatory component). Therefore, the OBSD is superior to the fast kurtogram for the bearing fault signature extraction under time-varying speed conditions.

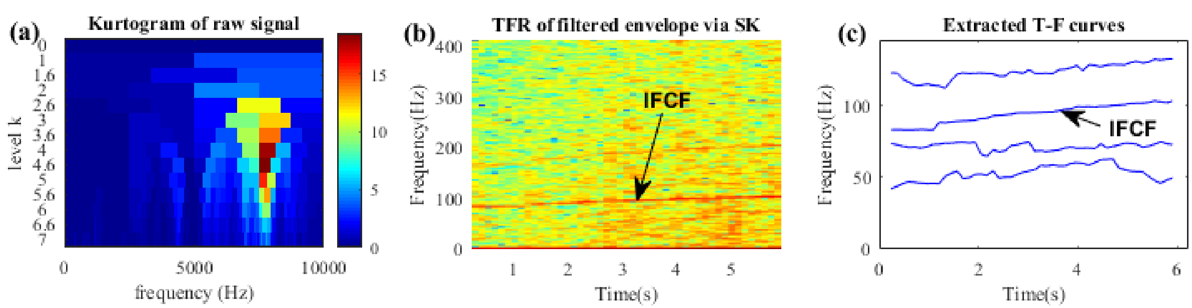

Fig. 6.21 Results of filtering via fast kurtogram for experiment 2. (a) Kurtogram of the raw signal, (b) TFR of filtered envelope via fast kurtogram, and (c) extracted T-F curves from (b).

6.6 Conclusions

In this chanpter, a method is proposed for bearing fault diagnosis under unknown time-varying speed conditions in the presence of an interference signal. The proposed method employs the OBSD method to extract the bearing fault signature from the contaminated signal. An MTFCE algorithm is then utilized to extract multiple T-F curves from the TFR of the processed signal. A challenge in the T-F curve-based methods is that the true ISRF may not appear in the T-F representation obtained via the OBSD, time-frequency curve extraction method or other technique. This could cause misleading diagnosis and severe machine damage. For this reason, an ISRF-search diagnosis strategy, as a crucial component of the proposed joint OBSD and MTFCE algorithm, is specifically designed to discover the true ISRF.

The performance of the proposed method is examined using four different types of simulated signals. The simulated signals are composed of a bearing fault-induced signal, an interference signal, and random noise. The simulated IFCF and IIF do not cross each other in simulation cases 1 and 2, and cross each other in simulation cases 3 and 4. Moreover, the ISRF is modulated by the bearing fault-induced signal in simulation cases 2 and 4, but not modulated in simulation cases 1 and 3. The results of all four simulations show that the proposed method can be used for effective bearing fault diagnosis under unknown time-varying speed conditions in the presence of interference. The performance of the OBSD method is also compared to the fast kurtogram method which is widely used for bearing vibration signal filtering. The results show that the fast kurtogram method is not able to obtain the expected optimal filter in the presence of strong interference signals. The comparison indicates that the OBSD method is superior for bearing fault signature extraction under time-varying speed conditions with the influence of interferences.

The effectiveness of the proposed method is also validated by signals collected from a test rig operated under time-varying speed conditions. The collected data contains not only the bearing vibrations signal but also the gear meshing signal, which is considered as an interference signal for bearing fault diagnosis. Signals are collected from two experiments conducted under different conditions. In the first experiment, the shaft and gearbox are driven by the same motor while in the second experiment, they are driven by two different motors. It is demonstrated that the fault can be effectively diagnosed in both experiments using the proposed method. The OBSD-based method obtains superior results in comparison to those obtained via the fast kurtogram.

Chapter 7 Short-time kurtogram for bearing fault feature extraction under time-varying speed conditions

This chapter addresses objective 5. A new method was proposed for bearing fault feature/signature extraction under time-varying speed conditions. Inspired by the STFT, the new method extends the kurtogram which is used for constant speed case to time-varying speed case by windowing the signal.

A portion of the content of this chapter has been accepted for presentation and publication by the *ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*.

7.1 Abstract

The kurtogram is a spectral analysis tool used to detect non-stationarities in a signal. It has been shown to be effective for locating the optimal filter for bearing fault feature extraction from the blurred vibration signal, since the transients of the bearing fault-induced signal are regarded as non-stationary. However, the effectiveness of the kurtogram is diminished when the signal is collected from a bearing operating under time-varying speed conditions. There is a need to improve the performance of the kurtogram under time-varying speed conditions. In this chapter, a short-time kurtogram method is proposed for bearing fault feature extraction under time-varying speed conditions. The performance of the short-time kurtogram is examined with simulated signals and experimental data. The results demonstrate that the short-time kurtogram can effectively extract the bearing fault features under time-varying speed conditions.

7.2 Introduction

Vibration signal processing is one of the most commonly used technique to detect bearing faults ([7], [9], [58]). Defects in the bearing induced impulse responses at a specific frequency called the Fault Characteristic Frequency (FCF), which is proportional to the bearing rotational frequency [9]. Hence, faults can be detected by observing the FCF and its harmonics in the frequency domain via signal processing. However, the collected signal is often smeared by random noise and interference signals transmitted from other sources. Therefore, extracting the bearing fault features from the contaminated signal is an important step for bearing fault detection and fault diagnosis.

The band-pass filter is one of the most efficient ways to filter a bearing signal for bearing fault detection ([113], [114]). Although it is known that a band-pass filter is one of the most basic tools for pre-filtering, the optimal center frequency and optimal bandwidth of the filter need to be determined. Since the bearing fault-induced signal can be considered as an amplitude modulated signal in which the resonance frequency is the frequency of the carrier signal, the resonance frequency is expected to be the center frequency of the optimal filter [113]. Additionally, the FCF and its multiples show sidebands around the resonance frequency in the spectrum of a bearing fault-induced signal [114]. Therefore, the optimal bandwidth should cover enough multiples of the FCF.

To locate the optimal band-pass filter, different methods have been proposed ([25], [37], [47]). The essence of these methods is to decompose/filter the signal via different filters, and then find the optimal filter by testing the decomposed/filtered signal. The wavelet transform is frequently implemented to decompose the signal. In this case, the optimal filter is dependent on wavelet-related parameters. A smoothness index was proposed to select the wavelet parameters and the parameters leading to the result with minimum smoothness index was considered to be the optimal parameters [25]. Additionally, the minimum Shannon entropy criterion was used to select the wavelet parameters [115]. Kurtosis has also been used to select the wavelet parameters for bearing fault feature extraction [37].

Among optimal band-pass filtering methods, Spectral Kurtosis (SK) based band-pass filtering is widely used to filter the bearing vibration signal for bearing fault detection. The concept of SK was first time introduced as the kurtosis in the frequency domain and calculated based on the frequency spectrum of a short segment of a signal [46]. It was found that the SK can be used to indicate the non-stationary transients in a signal [46]. The SK was also proposed based on the result of a uniform filter bank, and the concept of the kurtogram was introduced [24]. The uniform filter bank is also called a Short-Time Fourier Transform (STFT) based filter bank. Then, the proposed SK based filtering, i.e. kurtogram, was applied to filter the vibration signal for bearing fault detection [23]. To improve the computational efficiency, a fast computational kurtogram, called fast-kurtogram, was proposed based on the STFT filter bank [112]. Additionally, To enhance the accuracy of the kurtogram, minimum entropy deconvolution was implemented with the fast kurtogram [116]. Furthermore, an adaptive SK method was proposed by applying windows with non-uniform length in the frequency domain [84] and the adaptive SK can be applied for multiple faults diagnosis [50]. Moreover, instead of using the STFT based filter bank, the wavelet pocket transform was used for the calculation of the kurtogram ([117], [118]). The kurtogram has been proven a powerful tool to filter the vibration signal for bearing fault detection in accordance with applications [24]. However, the application of the SK is based on the assumption that the bearing operates at a constant speed. In reality, bearings often operate under time-varying speed conditions, which may degrade the performance of the kurtogram. Therefore, it is critical to improve the performance of the kurtogram for bearing fault feature extraction under time-varying speed conditions.

In this chapter, a short-time kurtogram is proposed to extract the bearing fault features under time-varying speed conditions. Inspired by the STFT, the temporal signal is windowed into segments with a short length. Then, the fast-kurtogram is implemented to the windowed segments, which result in a series of kurtograms along the time span. The result is called a Time-Kurtogram Representation (TKR). The rest of the chapter is organized as follows. In section 7.3, the basics of the kurtogram, including the SK and the fast-kurtogram are introduced. Then, the short-time kurtogram is proposed in section 7.4, and a method to determine the "final" optimal filter is also proposed. Following in section 7.5, simulations are conducted to test the performance of the proposed short-time kurtogram. In section 7.6, experimental data is also used to validate the effectiveness of the short-time kurtogram. In section 7.7, the results of section 7.5 and section 7.6 are discussed. Finally, in section 7.8, conclusions are drawn based on the results and the discussion.

7.3 The kurtogram

7.3.1 Spectral Kurtosis (SK)

The concept of SK was introduced as the kurtosis in the frequency domain [46]. The kurtosis of a stochastic process X is defined as the fourth-order moment of the process:

$$kurt(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{E\left[\left(X-\mu\right)^4\right]}{\left(E\left[\left(X-\mu\right)^2\right]\right)^2} \tag{7.1}$$

where μ is the mean value of X and σ is the standard variance of X. The SK of a process $x(n)$ is defined as [23]

$$K(f) = \frac{\left\langle \left|H(n,f)\right|^4 \right\rangle}{\left\langle \left|H(n,f)\right|^2 \right\rangle^2} - 2 \tag{7.2}$$

where $\langle f(n) \rangle$ stands for the temporal average of the function $f(n)$, $H(n, f)$ is the complex envelope or complex demodulated signal $x(n)$ at frequency f such that $H(n, f)e^{j2\pi nf}dZ(f)$ is the output at time n of a infinitely narrow-band filter centered on frequency f, where $dZ(f)$ is an orthonormal spectral increment. The SK is then obtained as a function of frequency. It has been found that the SK can be used to indicate non-stationarity at a certain frequency or within a certain band of frequencies [23]. A higher value of SK means additional transients are present within that band. Therefore, the presence of transients or non-stationary information can be located in the frequency domain via the SK.

With the ability to indicate the presence of non-stationary transients, the SK can be implemented to find the optimal filter for bearing fault feature extraction [24]. Since the collected vibration signal is often smeared by noise and interferences, filtering the signal before using the envelope spectrum to diagnose bearing faults can effectively improve the accuracy of the result [24]. A bearing fault-induced signal can be considered as an amplitude modulated signal with the resonance frequency as the frequency of the carrier signal and the Fault Characteristic Frequency (FCF) as the fundamental frequency of the modulating signal [9]. Ideally, the resonance frequency is the main peak in the frequency spectrum of the bearing fault induced signal and the sidebands appear on either side of the resonance frequency at *n**FCF intervals. Thus, for the optimal filter, the center frequency should be the resonance frequency and the bandwidth should cover sufficient multiples of the FCF. Theoretically, due to the non-stationary character of the bearing fault-induced signal, the optimal filter can be obtained by locating the band that leads to the maximum SK.

To locate the optimal filter, the SK has to be calculated in different bands, i.e. the obtained SK should be a function of frequency and bandwidth. Thus, the calculation of the SK is combined with signal decomposition. A filter bank is normally implemented to decompose the signal into components within different frequency bands [119]. With the filter bank, the signal is filtered with filters of different center frequencies and bandwidths. Then, the SK is calculated for each filtered signal. Therefore, the result of the SK can be plotted as a 3D figure, i.e. the representation of SK in the frequency-bandwidth plane, and this is called the kurtogram. There are many different types of filter banks that can be applied to the signal. In [23], a uniform filter bank was used to obtain the complex envelopes of the signal. An interpretation of the filter-bank is shown in Fig. 7.1. The signal $x(n)$ is decomposed into components at different frequencies. The envelopes $h(n, f)$ are obtained by further down-sampling the decomposed component. It can be seen from Fig. 7.1 that the bandwidths of the filters are uniform. However, filters with different bandwidth should be tested to locate the optimal filter, which implies that a series of uniform filter banks need to be applied to the signal. Therefore, the computational cost of the kurtogram obtained via a uniform filter bank is relatively high.

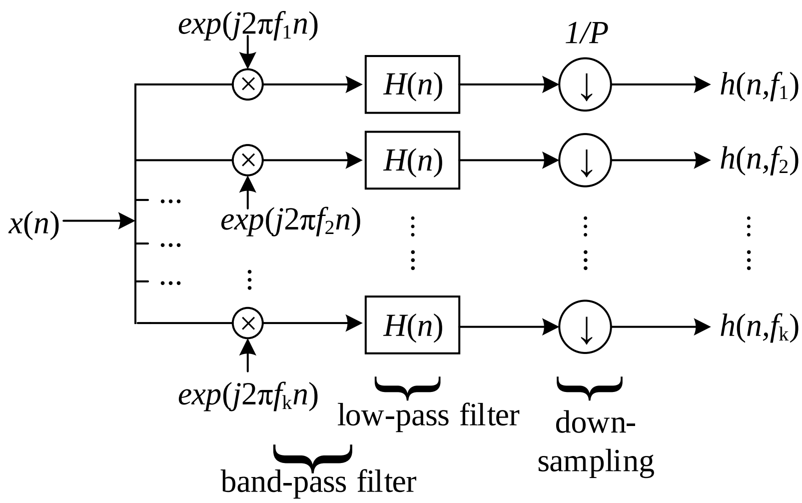

Fig. 7.1 Interpretation of a uniform filter bank

7.3.2 Fast-kurtogram

The computational cost of the kurtogram is dependent on the filter bank. To improve computational efficiency, a fast computational kurtogram, called the fast-kurtogram, was developed using an arborescent filter bank structure instead of a uniform filter bank [112]. The fast-kurtogram used a combination of binary tree and 1/3 tree structure to decompose the signal and then calculate the SK. The binary tree structure is shown in Fig. 7.2. The signal is decomposed by low-pass filters $H_0(n)$ and high-pass filters $H_1(n)$. The obtained coefficients $c_k^i(n)$ can be regarded as the complex envelope of the signal $x(n)$. The coefficients $c_k^i(n)$ can be considered to have been obtained via the i th filter, $i = 0, ..., 2^k - 1$, at the k th level with center frequency f_i and bandwidth $2(\Delta f)_k$. Assuming the sampling frequency is 1, the center frequency and the half bandwidth can be calculated from

$$f_i = \left(i + 2^{-1}\right) 2^{-k-1} \tag{7.3}$$

$$(\Delta f)_k = 2^{-k-1} \tag{7.4}$$

where i refers to the i th filter and k is the level of the filter bank.

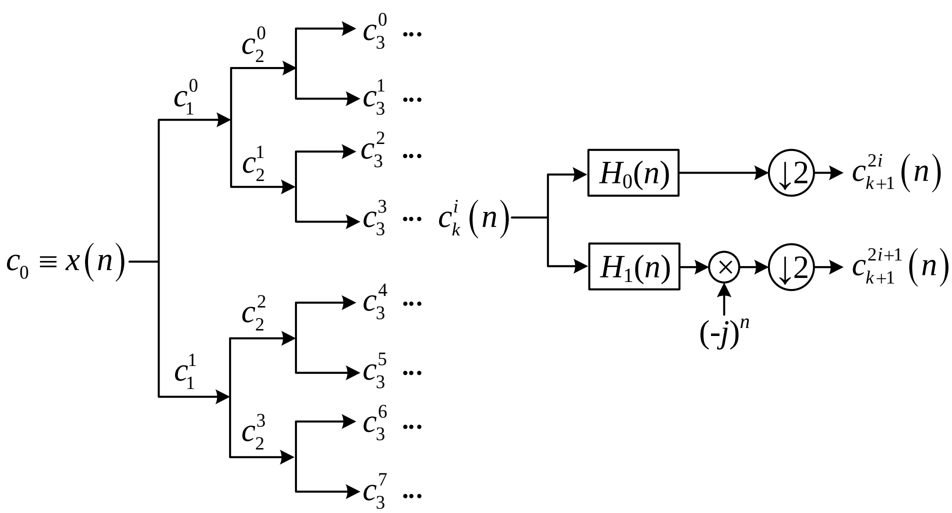

Fig. 7.2 A filter bank with binary tree structure

To enrich the binary tree, i.e. to apply more filters to the signal, the structure of the filter-bank is upgraded to a 1/3-binary tree. A scheme of the fast-kurtogram paving of the plane is shown in Fig. 7.3. Each sequence $c_k^i(n)$ obtained via the binary tree is further decomposed to three sub-sequences $c_{k+0.6}^{i+j}(n)$, j=0,1,2 with pass-band [0;1/6], [1/6;1/3], and [1/3;1/2], respectively. As shown in Fig. 7.3, the kurtogram is a frequency-bandwidth SK representation, i.e. a diagram of SK with respect to frequency and bandwidth $SK(f, \Delta f)$.

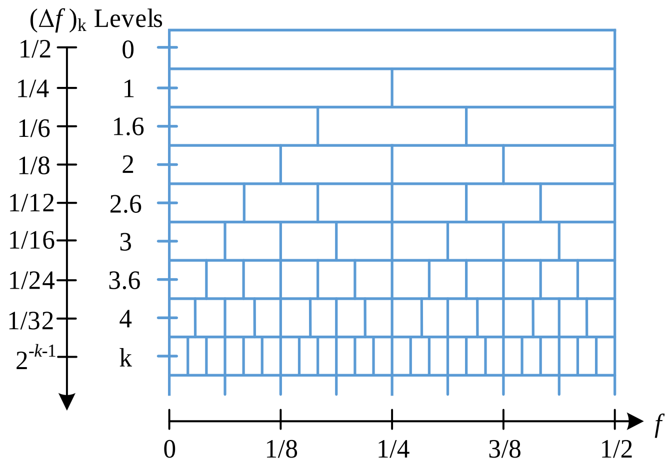

Fig. 7.3 Fast-kurtogram paving of the plane

The computational complexity of the kurtogram obtained via the 1/3-binary tree filter-bank is $O(KLN)$, where K is the total number of decomposition levels, L is the length of the prototype FIR filter $H(n)$, and N is the length of the signal. The complexity is effectively reduced compared to the kurtogram calculated via the uniform filter bank [112]. With high computational efficiency, the fast-kurtogram has been shown to be a useful tool to locate the optimal filter for bearing fault feature extraction [120]. However, the effectiveness of the fast-kurtogram is reduced if the bearing is running under time-varying speed conditions. The reason for this is that the expected optimal filter for the signal can also be time-varying since the changing rotational speed can increase the complexity of the vibration system and change the vibration signal. Therefore, it is important to improve the fast-kurtogram for time-varying speed cases.

7.4 Short-time kurtogram

To extend the application of the fast-kurtogram to bearing fault feature extraction under time-varying speed conditions, a short-time Kurtogram is proposed based on the short-time concept of the Short Time Fourier Transform (STFT). The fast-kurtogram is usually applied to the entire time span of a signal. However, if the bearing is operating under time-varying speed conditions, the result of the fast-kurtogram may not point to the expected optimal filter. Hence, the essence of the proposed short-time kurtogram is adding the time axis to the result of the fast-kurtogram, i.e. a Time-Kurtogram Representation (TKR), which is 4-dimensional. In this way, the result can indicate how the optimal filter changes along the time span, which in turn may improve the accuracy of the regular 3D kurtogram.

An interpretation of the short-time kurtogram is shown in Fig. 7.4. Similar to the STFT, the signal $x(n)$ of length N is first processed by a window $w_m(n_w)$ with length N_w ($N_w < N$), and the signal is separated into segments $x_m(n_w)$ by sliding the window. Then, the fast-kurtogram is applied to each segment to obtain the kurtogram $SK_m(f,\Delta f)$. Finally, all the obtained kurtograms generate a 4D plot, i.e. $SK(n_m, f, \Delta f)$, where $n_m = mN_w/2$. The total number of windows M can be determined from

$$M = \frac{N - N_w}{N_h} + 1 \tag{7.5}$$

where N_h is the hop size (also called step size) of the windows, i.e. the overlap of the windows is $N_w - N_h$. The computational complexity of the short-time kurtogram is $O(MKLN_w)$. If the overlap size of the windows is not too large, the complexity will not be much higher than the regular fast-kurtogram. Similar to the STFT, the result of the short-time kurtogram is related to the type of the implemented window, window size, and hop size/overlap of windows.

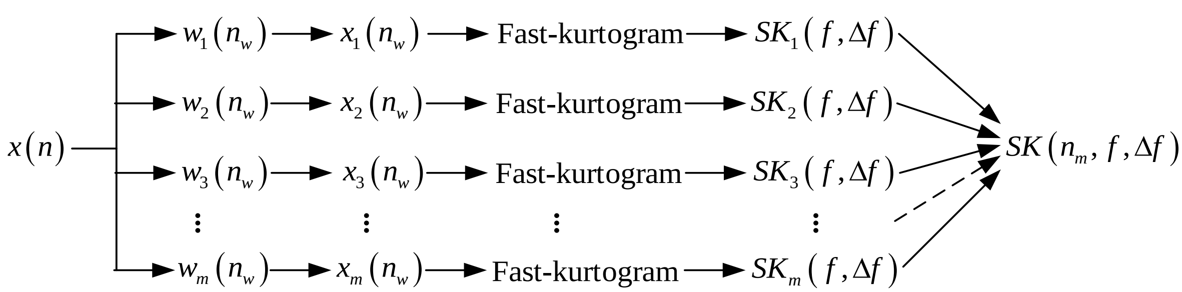

Fig. 7.4 Interpretation of the Short-Time-Kurtogram

7.4.1 Time-Kurtogram Representation (TKR)

An example of the TKR is shown in Fig. 7.5. The coordinate of the plot is comprised of the time axis, frequency axis, and level (bandwidth) axis. The figure is composed of the kurtogram obtained for each windowed segment. For each segment, the kurtogram is plotted at the center of the window in the time axis and an optimal filter can be located according to the kurtogram. Therefore, from the TKR, a series of optimal filters can be acquired for the signal along the time span.

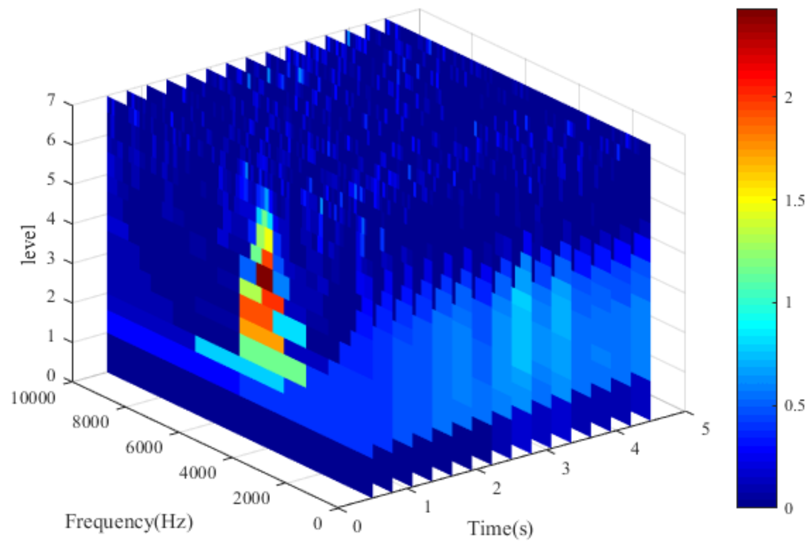

Fig. 7.5 An example of the TKR

7.4.2 Effects of the window

The result of the short-time kurtogram is related to the selection of the window parameters, including the window function $w(n)$, window length, and the hop size N_h (overlapping of the window). Additionally, the choice of window impacts the complexity of the short-time kurtogram.

The window function affects the result of the kurtogram. There are many different types of window functions, such as hamming window, Hanning window, rectangular window, etc. Each window has different properties when applied to the signal. However, it is found that the rectangular window should be chosen to conduct the short-time kurtogram since it performs better than other window functions for the analysis of transients [121]. As previously mentioned, the kurtogram locates the optimal filter by detecting the transients in the signal. Any window other than the rectangular window changes the amplitude of the transients which will in turn reduce the accuracy of the kurtogram. The rectangular window function is defined as [122]

$$w(n)=\begin{cases}1, & 0\le n\le N_w-1\\ 0, & otherwise\end{cases} \tag{7.6}$$

It should be pointed out that the discussion here focuses on the effects of the window function on the result of the kurtogram, not on the effect on the signal reconstruction. Also, the window function has no effect on the computational complexity of the short-time kurtogram.

The window length N_w is related to the frequency resolution of the kurtogram. It is known that the length of the windowed signal is related to the frequency resolution of its frequency spectrum, i.e. a longer signal leads to a higher frequency resolution [122]. Therefore, if the window length is shorter, the frequency resolution is poorer, and the narrowest bandwidth available in the kurtogram is narrower. To ensure that the bandwidth of interest is within the appropriate scope, the window length should be chosen to be long enough. Moreover, a longer window length can make the windowed segment include more transients, which can improve the result of the kurtogram. However, for a bearing vibration signal collected under time-varying speed conditions, if the window length is too long, the time-varying speed may decrease the accuracy of the kurtogram. Additionally, the complexity of the short-time kurtogram is dependent on the window length. It is given that the complexity of the short-time kurtogram is $O(MKLN_w)$ and N_w is also related to *M*. According to the complexity and Eq. (7.5), we can show that if $N_w>(N-N_h)/2$, the complexity increases with the increase of N_w, otherwise, the complexity decreases. Normally, the window length satisfies $N_w>(N-N_h)/2$. Therefore, the complexity of the short-time kurtogram increases with the increase in window length. It then follows that the selection of the window length should consider the bandwidth size of interest, the rate of change of the time-varying speed, as well as computational complexity.

The hop size N_h of the windows is related to the time resolution of the result of the short-time kurtogram. A smaller hop size results in a better time resolution, i.e. the time interval between two kurtograms is shorter. For a bearing signal collected under time-varying speed conditions, the hop size should be small enough to show more kurtograms along the time span. However, according to Eq. (7.5), a smaller hop size increases the total number of the kurtograms, which leads to higher computational complexity. Therefore, the selection of the hop size should consider the rate of change of time-varying speed, as well as computational complexity.

7.4.3 Signal filtering via the result of short-time kurtogram

With the proposed short-time kurtogram, the TKR and a series of the optimal filters can be obtained along the time span. However, it is also important to properly make use of the result of the short-time kurtogram to achieve bearing fault feature extraction under time-varying speed conditions. For the kurtogram, the optimal filter is determined by locating the band with the maximum SK, since the SK indicates the level of the non-stationarity in a signal. To extend this idea to the short-time kurtogram, it is proposed to choose the “final” optimal filter from the optimal filters by selecting the one with the maximum SK value.

A flowchart of bearing signal filtering using the result of the short-time kurtogram is shown in Fig. 7.6. From the result of the short-time kurtogram, the parameters of the optimal filter for each segmented signal can be obtained, i.e. the center frequency $f_c\left(n_m\right)$ and half bandwidth $\Delta f\left(n_m\right)$. Then, the optimal filter with the maximum SK value is chosen as the final optimal filter and the corresponding parameters f_c and Δf are also obtained. Correspondingly, the final optimal filter $h\left(n\right)$ can be designed. Finally, the signal $x\left(n\right)$ is filtered and the result is $y\left(n\right)$.

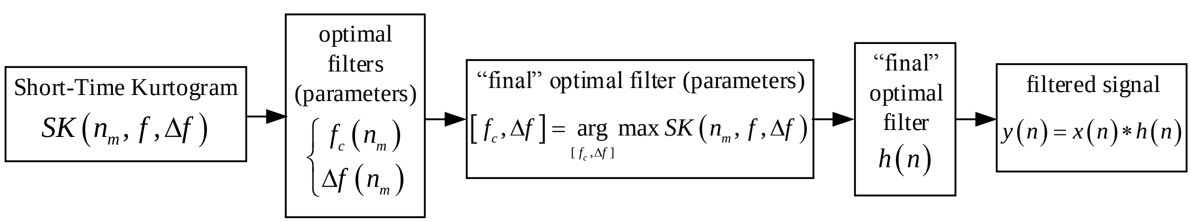

Fig. 7.6 Flowchart of signal filtering using the result of short-time kurtogram

7.5 Simulations

To test the performance of the proposed short-time kurtogram for bearing fault feature extraction under time-varying speed conditions, the short-time kurtogram is applied to simulated signals. The signal is simulated as a mixture of a bearing fault-induced signal, interferences and random noise.

The bearing fault-induced signal can be considered as a series of impulse responses which repeat periodically at the FCF [32]. For a bearing operating under time-varying speed conditions, the signal can be simulated as [81]

$$\begin{aligned} x_B(t) &= \left[1 + A(t)\right]\sum_{p=1}^{P} s_p(t) \\ &= \left[1 + \alpha\cos\left(2\pi f_r\right)\right]\sum_{p=1}^{P} L_p e^{-\beta\left(t-t_p\right)} \sin\left[\omega_r\left(t-t_p\right)+\phi_p\right] u\left(t-t_p\right) \end{aligned} \tag{7.7}$$

where $A(t)=\alpha\cos\left(2\pi f_r t\right)$ refers to the modulated waveform, f_r is the rotational frequency, α is the amplitude of the modulation ($0<\alpha<1$ for inner race fault and $\alpha=0$ for outer race fault [9]), P is the number of impulse responses which is determined by the signal length and FCF, $s_p(t)$ refers to the pth impulse response, L_p is the amplitude of the pth impulse response, β is the coefficient related to damping, ω_r is the excited resonance frequency or damped frequency of the vibration system, ϕ_p is the phase of the pth impulse response, and $u(t)$ is unit step function. In the previous equation, t_p is the occurrence time of the pth impulse response which is calculated as

$$\begin{cases} t_1 = \left(1+\delta_1\right)\left[1/f_{cf}(0)\right] \\ t_p = \left(1+\delta_p\right)\left[1/f_{cf}(0) + 1/f_{cf}(t_1) + \ldots + 1/f_{cf}(t_{p-1})\right] \qquad p = 2,3,\ldots,P \end{cases} \tag{7.8}$$

where δ_p is the random slippage ratio with average varying between 0.01 and 0.02, $f_{cf}(t)$ denotes the Instantaneous Fault Characteristic Frequency (IFCF), and the time interval between the (p-1)th impulse response and the pth impulse response is $\left(1+\delta\right)/f_{cf}\left(t_{p-1}\right)$.

The interference can be simulated as a sum of sinusoidal functions at the frequency of the interference and its harmonics, given as [81]

$$I(t) = \sum_{n_i=1}^{N_i} B_{n_i}\sin\left(2\pi n_i f_I t\right) \tag{7.9}$$

where N_i is the total number of sinusoidal harmonics, B_{n_i} is the amplitude, and f_I is the time-varying frequency of the interference, also called Instantaneous Interference Frequency (IIF).

The interference signal is he vibration signal transmitted from other sources in the vibration system, which is possibly more than one source. It is indicated in [109] that the vibration signal also comprises the interference with the Instantaneous Shaft Rotational Frequency (ISRF) as the fundamental frequency caused by misalignment, eccentricity or imbalance. Therefore, the interference signal can be simulated as

$$x_I(t) = I_1(t) + I_2(t) = \sum_{n_{1i}=1}^{N_1} B_{1n_i} \sin\left(2\pi n_{1i} f_I t\right) + \sum_{n_{2i}=1}^{N_2} B_{2n_{2i}} \sin\left(2\pi n_{2i} f_r t\right) \tag{7.10}$$

where $I_1(t)$ is the same as $I(t)$ given in Eq. (7.9) and $I_2(t)$ is the interference with the ISRF as the fundamental frequency. Finally, the faulty bearing signal contaminated by interferences and noise is simulated with

$$x(t) = x_B(t) + x_I(t) + n(t) \tag{7.11}$$

where $n(t)$ refers to random noise with intensity specified by the SNR.

To examine the performance of the short-time kurtogram with different types of faults, two simulations are conducted. In the first case, the bearing fault-induced signal is simulated with amplitude modulation, i.e. the fault is inner race fault. In the other case, the bearing fault-induced signal is simulated without amplitude modulation, which is similar to the case of outer race fault. In both cases, the rotational frequency is speeding up.

7.5.1 Case 1: simulated signal with amplitude modulation

A signal is simulated with values f_r =(10t+20) Hz (ISRF), α =0.9, L_p =1, β =1000, ω_r =4000*2π rad/s, ϕ_p =0, and where δ_p is a random number with a mean of 0.01, f_{cf} =3.5 f_r (IFCF), N_1=3, f_I =(15t+35)Hz (IIF), B_1=[1, 0.5, 0.05], N_2 =2, B_2 =[0.6, 0.3], and SNR=-5 dB. The sampling rate is 20kHz and the sampling time is 5s. The simulated signal is shown in Fig. 7.7(a). Normally, for bearings operating under constant speed condition, the faults are diagnosed by detecting the FCF and its harmonics in the frequency spectrum of the envelope signal [9]. The envelope of a signal $x(t)$ can be obtained via the Hilbert transform [104]

$$H\left[x(t)\right] = \frac{1}{\pi}\int_{-\infty}^{\infty} x(\tau)\frac{1}{t-\tau}\,\mathrm{d}\tau \tag{7.12}$$

Since the bearing is operating under time-varying speed conditions, the FCF is also time-varying, which means that bearing faults cannot be simply detected in the frequency domain. The STFT can be used to obtain the Time-Frequency-Representation (TFR) of the envelope signal, in which the IFCF and its harmonics can be detected. The STFT is defined as [122]

$$X(\tau, f) = \int_{-\infty}^{+\infty} x(t) w(t-\tau) \mathrm{e}^{-\mathrm{j}2\pi f t} \mathrm{d}t \tag{7.13}$$

where $w()$ is the window function, τ is the time variable, and f is the frequency variable. By implementing the Hilbert transform and the STFT, the TFR of the envelope signal (denoted as envelope TFR) is obtained and shown in Fig. 7.7(b). In the obtained envelope TFR of the simulated signal, the time-frequency curves are messy. The IFCF and its harmonics cannot be clearly observed in Fig. 7.7(b). Obviously, additional signal filtering is needed to extract the bearing fault features from the contaminated original signal.

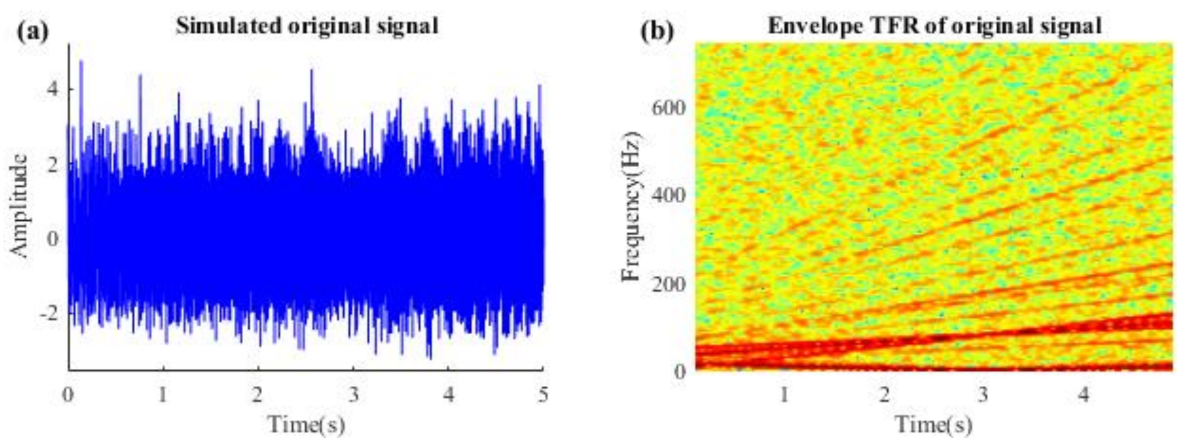

Fig. 7.7 (a) simulated signal of case 1 and (b) its envelope TFR

The regular kurtogram is obtained by applying the fast-kurtogram to the entire time span of the simulated signal, as shown in Fig. 7.8(a). From the kurtogram, an optimal filter is located at level 7 with center frequency 273Hz (the bandwidth is about 78Hz). With this optimal filter, the signal is filtered and the envelope TFR is shown in Fig. 7.8(b). In the envelope TFR of the filtered signal, the time-frequency curves are still messy. The IFCF and its harmonics cannot be observed in Fig. 7.8(b). The reason is that the optimal filter is not properly obtained. It is mentioned that the center frequency of the expected optimal filter should be close to the resonance frequency ω_r, which is set as 4000Hz in the simulation. However, the center frequency of the obtained optimal frequency is 273Hz, which is too low. From the results shown in Fig. 7.8(a) and Fig. 7.8(b), it can be seen that the fast-kurtogram is not useful for bearing fault feature extraction under time-varying speed conditions in this simulation.

By applying the short-time kurtogram to the simulated signal, the TKR is obtained and shown in Fig. 7.8(c). The window length is set as N_w =10000 samples (0.5s) and the hop size is set as N_h =4736 samples (0.24s). The total number of the windows is then calculated as M =20. From the TKR, 20 optimal filters are located for the signal. The center frequency f_c and the cut-off frequencies ($f_c - \Delta f$ and $f_c + \Delta f$) of the filters are shown in Fig. 7.8(d). The corresponding SK values of the 20 optimal filters are also shown in Fig. 7.8(e). The “final” optimal filter is picked as the one obtained from the first windowed segment since the SK value is the maximum in Fig. 7.8(e). The center frequency of the filter is 4062Hz and the bandwidth is 622Hz. By applying the corresponding filter to the signal, the envelope TFR is obtained and shown in Fig. 7.8(f). The ISRF, the IFCF, and its harmonics can be clearly observed in Fig. 7.8(f) which reveals that the bearing fault features are extracted from the simulated signal. Additionally, the interferences are removed from the signal since the IIF and its harmonics cannot be observed in Fig. 7.8(f). The results show that the short-time kurtogram can be used to effectively extract the bearing fault features from the contaminated signal under time-varying speed conditions.

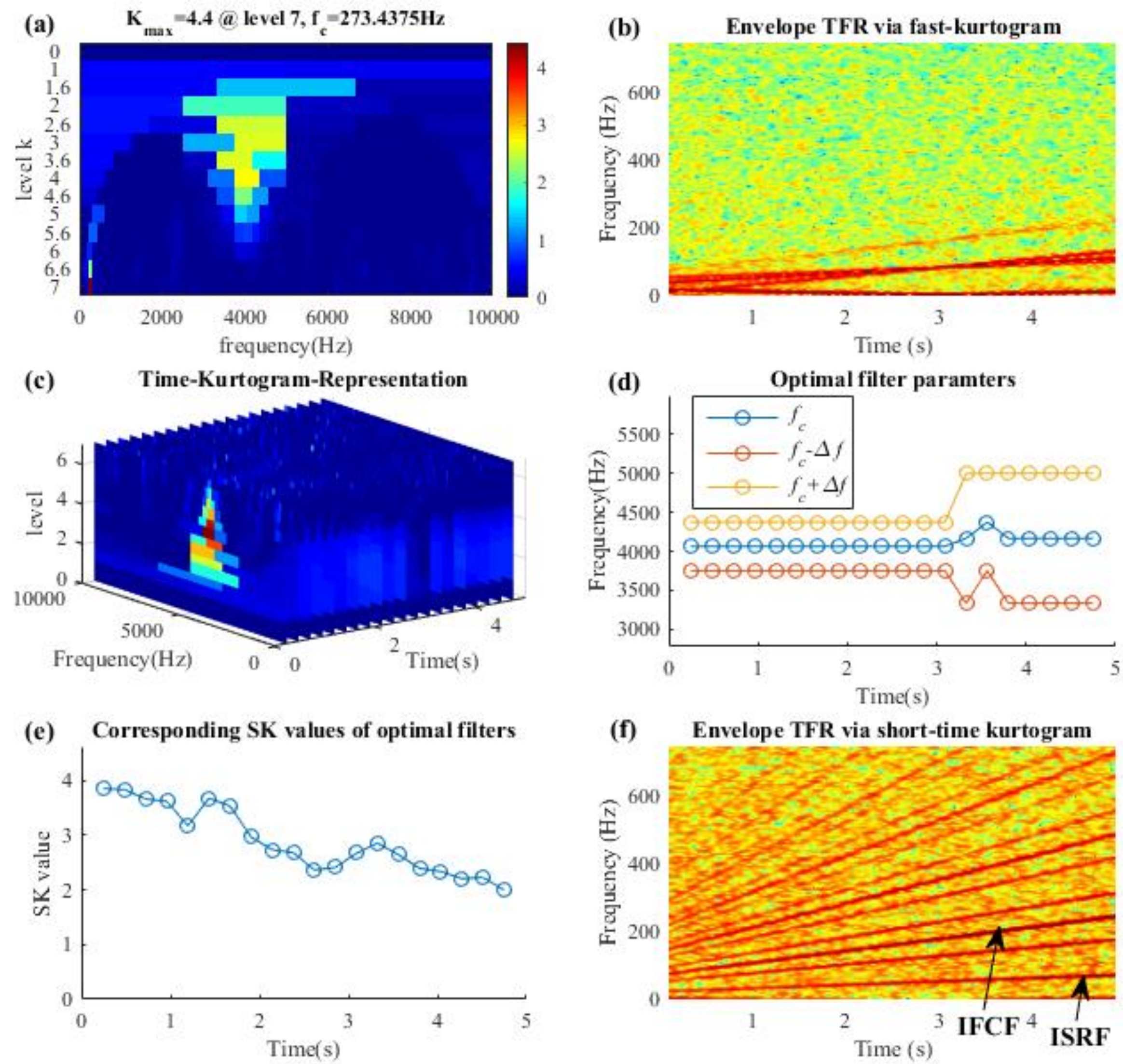

Fig. 7.8 Results of case 1: (a) kurtogram obtained via the fast-kurtogram, (b) envelope TFR of the filtered signal via the fast-kurtogram, (c) TKR obtained via the short-time kurtogram, (d) parameters of a series optimal filters, (e) corresponding SK values of the optimal filters, and (f) envelope TFR of the filtered signal via the short-time kurtogram.

Comparing the optimal filter obtained via the single fast-kurtogram (shown in Fig. 7.8(b)) to the optimal filters obtained via the short-time kurtogram (shown in Fig. 7.8(d)), the optimal filters obtained from the windowed segments are more accurate than the one obtained directly from the full-length signal. All the optimal filters shown in Fig. 7.8(d) have the center frequency around 4000Hz which is where the optimal filter was expected to be. The rotational frequency increases 5Hz and the interference frequency f_I increases 7.5Hz in each segment. However, the center frequency of the optimal filter shown in Fig. 7.8(b) from the single fast-kurtogram is 273Hz, which is too low compared to 4000Hz. The rotational frequency increases 50Hz and the interference frequency increases 75Hz in the full-length signal. This comparison reveals that the fast-kurtogram is still effective for bearing fault feature extraction when the rotational frequency and the interference frequency do not vary too much, however, it may not be effective if the frequencies changes significantly.

It can be also observed from Fig. 7.8(d) that the bandwidth of the optimal filter increases with the increase of the rotational frequency. It is mentioned that the bandwidth of the expected optimal filter should cover enough multiples of the FCF. Since the rotational frequency is simulated as increasing along the time span, the IFCF also increases, which means the bandwidth of the optimal filter should become wider. The result shown in Fig. 7.8(d) coincides with the expected case. Therefore, in view of this, the effectiveness of the short-time kurtogram is validated again.

7.5.2 Case 2: simulated signal without amplitude modulation

Here, a simulation is performed to validate the effectiveness of the short-time kurtogram for the case of shaft rotational speed with a slow down/speed up pattern.

A new signal is simulated, shown in Fig. 7.9(a). Different from the simulation case 1, the signal is simulated without amplitude mosulation, i.e. $\alpha = 0$. The other settings are kept the same as in case 1, which are f_r =(10t+20) Hz (ISRF), L_p =1, β =1000, ω_r =4000*2π rad/s, ϕ_p =0, δ_p is a random number with a mean of 0.01, f_{cf} =3.5 f_r (IFCF), N_1=3, f_I =(15t+35)Hz (IIF), B_1=[1, 0.5, 0.05], N_2 =2, B_2 =[0.6, 0.3], and SNR=-5 dB. The sampling rate is 20kHz and the sampling time is 5s. The envelope TFR of the simulated signal is obtained via Hilbert transform and STFT, as shown in Fig. 7.9(b). The envelope TFR is dominated by the IIF. The IFCF and its harmonic are smeared by the interferences.

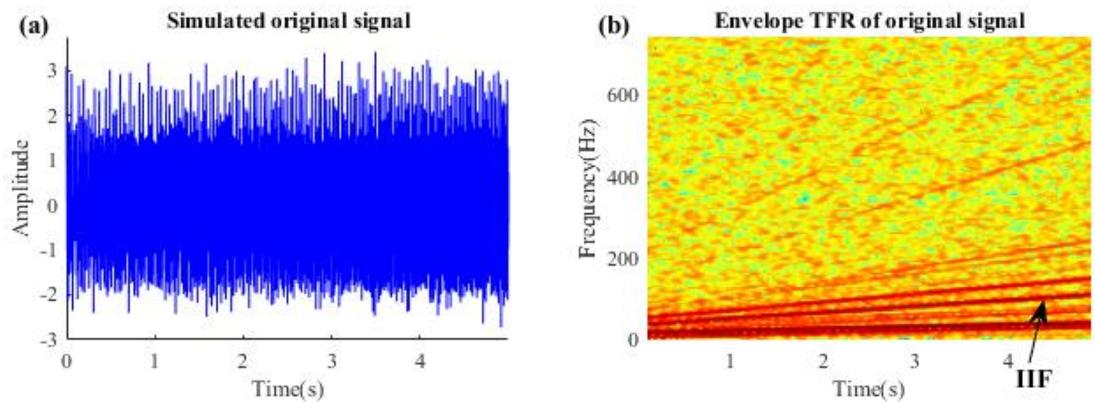

Fig. 7.9 (a) simulated signal of case 2 and (b) its envelope TFR

Applying the fast-kurtogram to the full-length simulated signal, a single kurtogram is obtained and shown in Fig. 7.10(a). According to the single kurtogram, the optimal filter is located at level 7 (bandwidth 78Hz)with the center frequency 273Hz. Implementing the located optimal filter, the envelope TFR of the filtered signal is shown in Fig. 7.10(b). Compared to the envelope TFR of the original signal, the harmonics of the IFCF are completely removed in the envelope TFR of the filtered signal. The envelope TFR is also dominated by the IIF. The result indicates that the optimal filter obtained via the single kurtogram shown in Fig. 7.10(a) is incorrect. This can be also noticed from the center frequency (273Hz) of the optimal filter, which is far from the expected center frequency of 4000Hz. Therefore, the single fast-kurtogram cannot handle the time-varying speed in this simulated case.

The TKR of the simulated signal is obtained via the short-time kurtogram, shown in Fig. 7.10(c). The parameters are set the same as case 1: window length N_w =10000 samples (0.5s) and hop size N_h =4736 samples (0.24s). The total number of windows is then calculated as $M = 20$. From the kurtograms of the segmented signal, 20 optimal filters are obtained. The center frequency f_c and the cut-off frequencies $f_c \pm \Delta f$ of each optimal filter are shown in Fig. 7.10(d). Additionally, the corresponding SK values of the optimal filters are shown in Fig. 7.10(e). From Fig. 7.10(e), the 2nd filter contains the maximum SK value. Thus, the 2nd filter is selected as the “final” optimal filter. The center frequency of the filter is 4062Hz and the bandwidth is 622Hz. Applying this filter to the simulated signal, the envelope TFR of the filtered signal is obtained, shown in Fig. 7.10(f). From Fig. 7.10(f), the IFCF and its harmonic can be clearly observed, which indicates that the bearing fault features are effectively extracted from the contaminated signal. Also, the interferences are completely removed from the original signal. Therefore, the result shows that the short-time kurtogram can be effectively implemented to extract the bearing fault features under time-varying speed conditions.

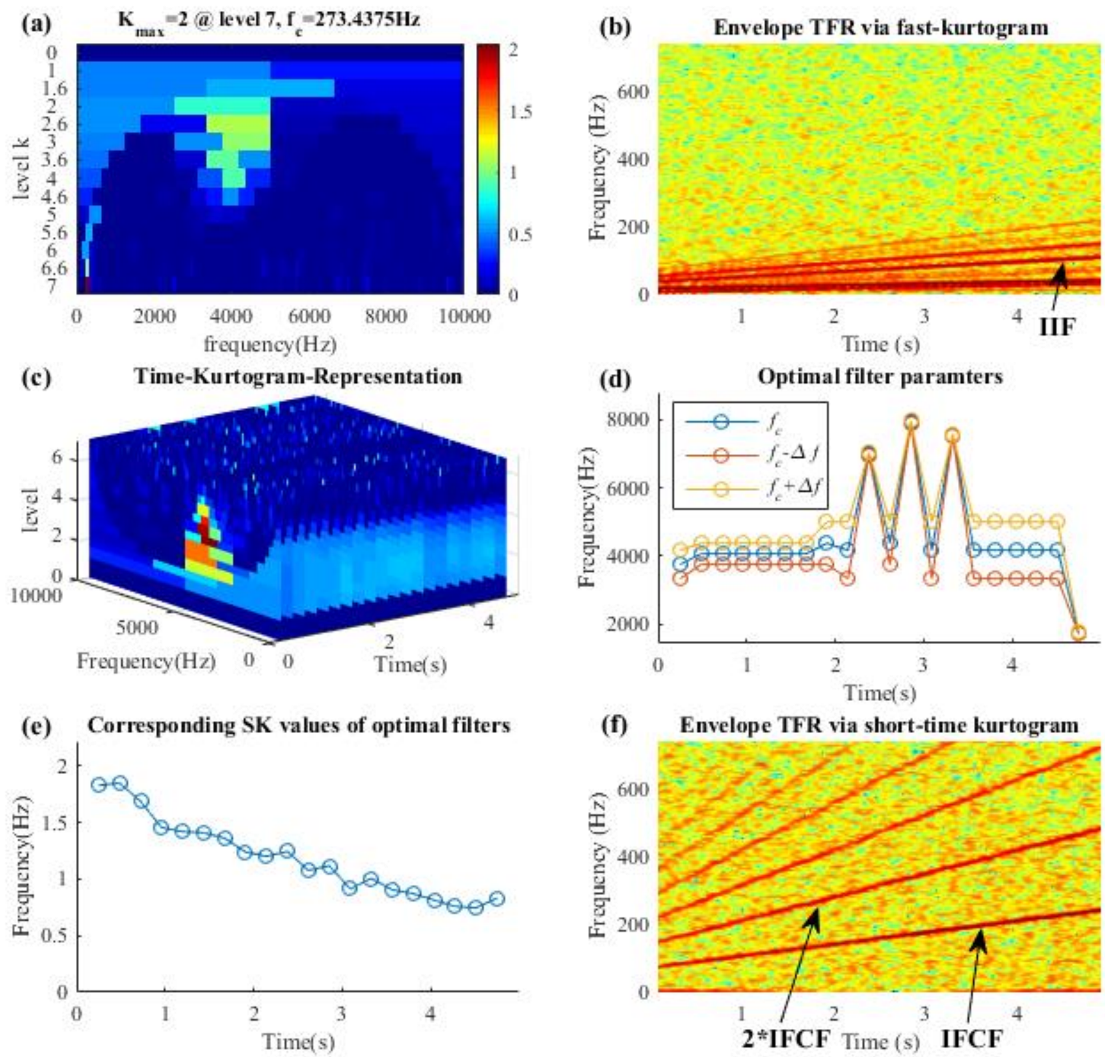

Fig. 7.10 Results of case 2: (a) kurtogram obtained via the fast-kurtogram, (b) envelope TFR of the filtered signal via the fast-kurtogram, (c) TKR obtained via the short-time kurtogram, (d) parameters of a series optimal filters, (e) corresponding SK values of the optimal filters, and (f) envelope TFR of the filtered signal via the short-time kurtogram.

Similarly, by comparing the result of the single fast-kurtogram to the result of the short-time kurtogram, it reveals that the fast-kurtogram may not be useful if the time-varying frequencies change too much. However, by segmenting the signal to reduce the change of the rotational frequency and the interference frequency, the fast-kurtogram can still be used.

Comparing the result shown in Fig. 7.10(d) to the result shown in Fig. 7.8(d), not all the obtained optimal filters shown in Fig. 7.10(d) have the center frequencies around 4000Hz. Three of them have the center frequency between 7000Hz to 8000Hz, which is much higher than 4000Hz. However, by choosing the optimal filter with the maximum SK value, the final optimal filter is selected as the one with the expected center frequency. This indicates that even not all the segmented signals get proper optimal filters, the result of the short-time kurtogram can still be effective.

7.6 Experiments

To further test the performance of the short-time kurtogram, experimental data are collected from a vibration system which includes a faulty bearing. The set-up of the experiment is shown in Fig. 7.11. A 1-inch shaft is supported by two bearings, one is healthy and the other one is faulty. The shaft is driven by a motor and the rotational speed is controlled by an AC driver. An accelerometer is installed on top of the housing of the faulty bearing to collect the vibration signal. An encoder is used to measure the rotational speed. Additionally, a gearbox is connected to the shaft. Therefore, the collected vibration signal contains not only the bearing fault-induced signal but also the gear meshing (interference) signal. The structural parameters of the bearings and gears used in the experiments are given in Table 7.1. According to the parameters of the bearings, the FCF of the bearing outer race, i.e. Ball-Pass Frequency of Outer-race (BPFO), is equal to the product of the FCF coefficient (3.57) and the shaft rotational frequency f_r [9], i.e. BPFO=3.57 f_r. Similarly, the FCF coefficient of the bearing inner race is 5.43, i.e. Ball-Pass Frequency of Inner-race (BPFI) = 5.43 f_r. In addition, the gear meshing frequency is calculated as the shaft rotational frequency divided by the diameter ratio of the sheaves and then multiplied by the number of the pinion teeth, i.e. $f_{meshing} = \left(18/2.6\right) f_r = 6.92 f_r$.

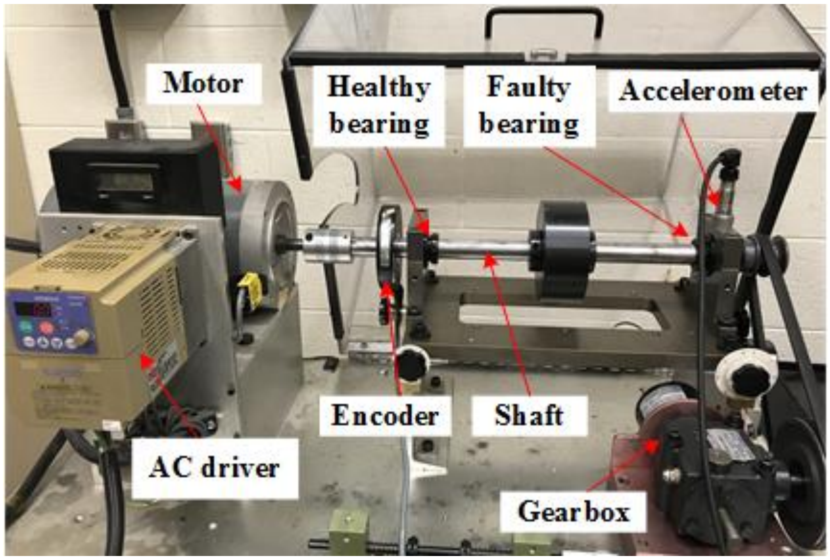

Fig. 7.11 Experimental set-up

Table 7.1 Parameters of bearings and gears

Bearing type	Pitch diameter	Ball diameter	Number of balls	BPFO	BPFI
ER16K	38.52mm	7.94mm	9	$3.57f_r$	$5.43f_r$
Diameter ratio of sheaves		Number of teeth		Gear meshing frequency	
1:2.6		18		$6.92f_r$	

7.6.1 Experiment 1: bearing with outer race fault

In the first experiment, the faulty bearing has a local defect on the outer race. The signal is sampled with sampling frequency 20kHz and the signal length is 10s. The collected vibration signal is shown in Fig. 7.12(a) and the measured shaft rotational speed is shown in Fig. 7.12(b). The shaft rotational speed is increased almost linearly from 15Hz to 25Hz. The TFR of the collected signal is obtained via the STFT, shown in Fig. 7.12(c). The TFR is dominated by the second harmonic of the gear meshing frequency and higher order of harmonics. The IFCF of the outer race fault cannot be clearly observed in Fig. 7.12(c). Applying the Hilbert transform and the STFT to the collected signal, the envelope TFR of the original signal is obtained, shown in Fig. 7.12(d). The gear meshing frequency can be observed along the left half of the time span. However, almost no time-frequency curves in the envelope TFR can be seen clearly along the entire time span. Obviously, the signal needs to be further processed to extract the bearing fault features.

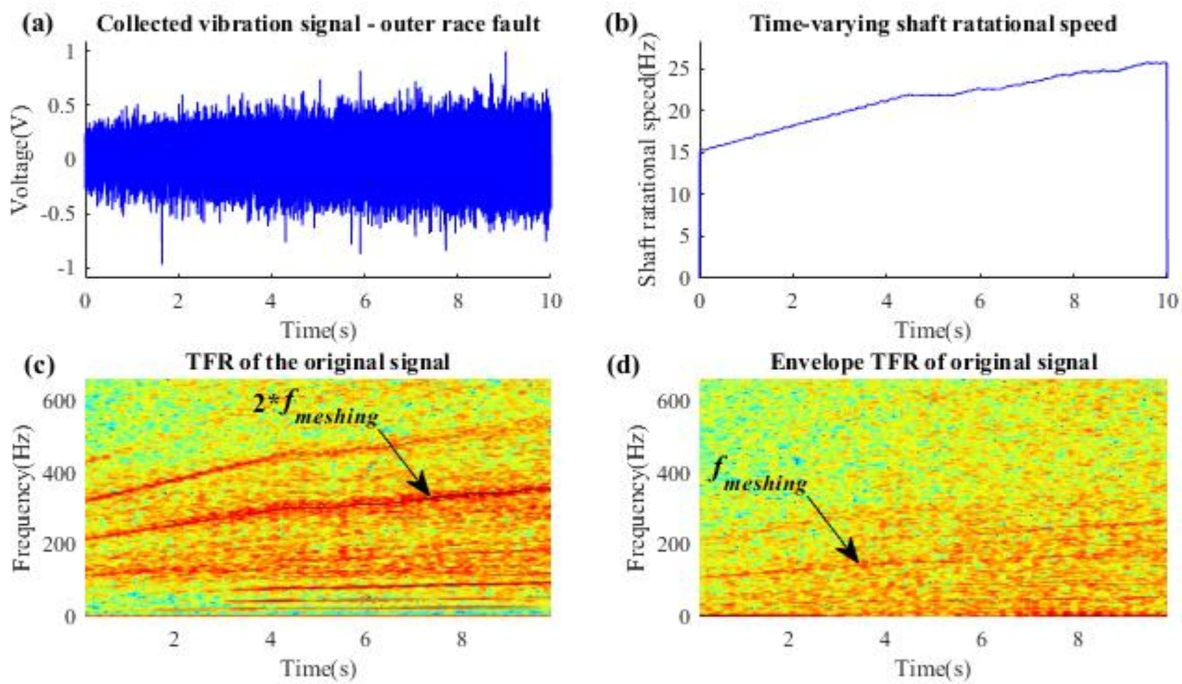

Fig. 7.12 (a) Collected vibration signal of experiment 1, (b) measured time-varying rotational speed, (c) TFR of the original signal, and (d) envelope TFR of the original signal.

The fast-kurtogram is applied to the collected full-length vibration signal and the obtained single kurtogram is shown in Fig. 7.13(a). According to the kurtogram, the optimal filter is located at level 5 (bandwidth 312Hz) with the center frequency 781Hz. The signal is then filtered by the optimal filter and the envelope TFR of the filtered signal is shown in Fig. 7.13(b). The obtained envelope TFR is similar to the envelope TFR shown in Fig. 7.12(d). The result indicates that the bearing fault features are not properly extracted by implementing the fast-kurtogram with the full-length signal.

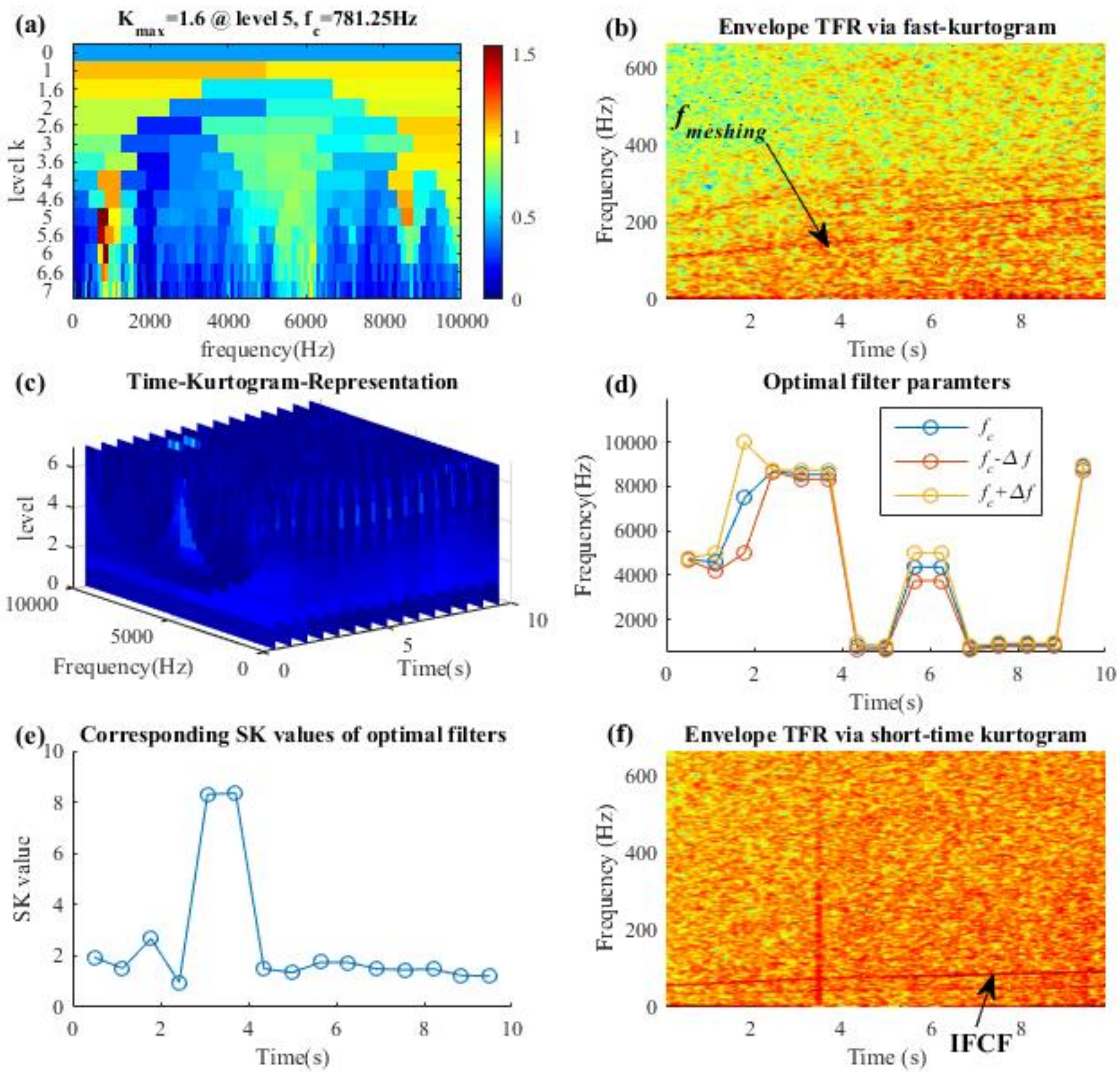

Fig. 7.13 (a) Kurtogram of the original signal obtained via the fast-kurtogram, (b) envelope TFR of the filtered signal via the fast-kurtogram, (c) TKR obtained via the short-time kurtogram, (d) parameters of a series optimal filters, (e) corresponding SK values of the optimal filters, and (f) envelope TFR of the filtered signal via the short-time kurtogram.

The short-time kurtogram is applied to the collected signal and the obtained TKR is shown in Fig. 7.13(c). The window length is set as 20000 samples (1s) and the hop size is set as 12857 samples (about 0.64s). The total number of the windows is 15. The optimal filters obtained for the 15 windowed segments are shown in Fig. 7.13(d). The corresponding SK values of the optimal filters are plotted in Fig. 7.13(e). The SK value of the 6th filter is the maximum. Therefore, the 6th optimal filter is applied to the original signal with a center frequency of 8334Hz and bandwidth 415Hz. The envelope TFR of the filtered signal is shown in Fig. 7.13(f). The FCF of the outer race fault can be clearly observed in Fig. 7.13(f). Additionally, the gearing meshing frequency is totally removed from the TFR which indicates that the gear meshing signal is filtered out from the collected signal. The result reveals that the short-time kurtogram can be effectively used to extract the bearing fault features for the bearing with outer race fault operating under time-varying speed conditions.

Comparing the result of the fast-kurtogram shown in Fig. 7.13(b) to the result of the short-time kurtogram shown in Fig. 7.13(f), the center frequency of the optimal filter obtained via the single fast-kurtogram (781Hz) is much lower than the center frequency (8334Hz) obtained via the short-time kurtogram. The total change of the shaft rotational frequency is 10Hz and the change of the gear meshing frequency (interference frequency) is approximately 69Hz which are larger than the changes in the windowed segment (1Hz for the rotational frequency and 6.9Hz for the gear meshing frequency). The comparison indicates that the fast-kurtogram can be still effective for application to bearing fault feature extraction under time-varying speed conditions if the signal is properly segmented.

7.6.2 Experiment 2: bearing with inner race fault

To further examine the performance of the short-time kurtogram, a second experiment is conducted by replacing the faulty bearing with another one vaing an inner race fault. The signal is sampled with sampling frequency 20kHz and the signal length is 10s. The collected vibration signal is shown in Fig. 7.14(a) and the measured shaft rotational speed is shown in Fig. 7.14(b). The shaft rotational speed decreases from 24Hz to 12Hz. The TFR of the collected vibration signal is obtained via STFT, as shown in Fig. 7.14(c). The TFR is dominated by the 2nd harmonic of the gear meshing frequency and higher order harmonics. The envelope TFR of the collected signal is obtained via Hilbert transform and the STFT, shown in Fig. 7.14(d). Similar to the envelope TFR shown in Fig. 7.12(d), the IFCF of the inner race fault cannot be observed in the envelope TFR, which implies that signal filtering is needed to extract the bearing fault features from the contaminated vibration signal.

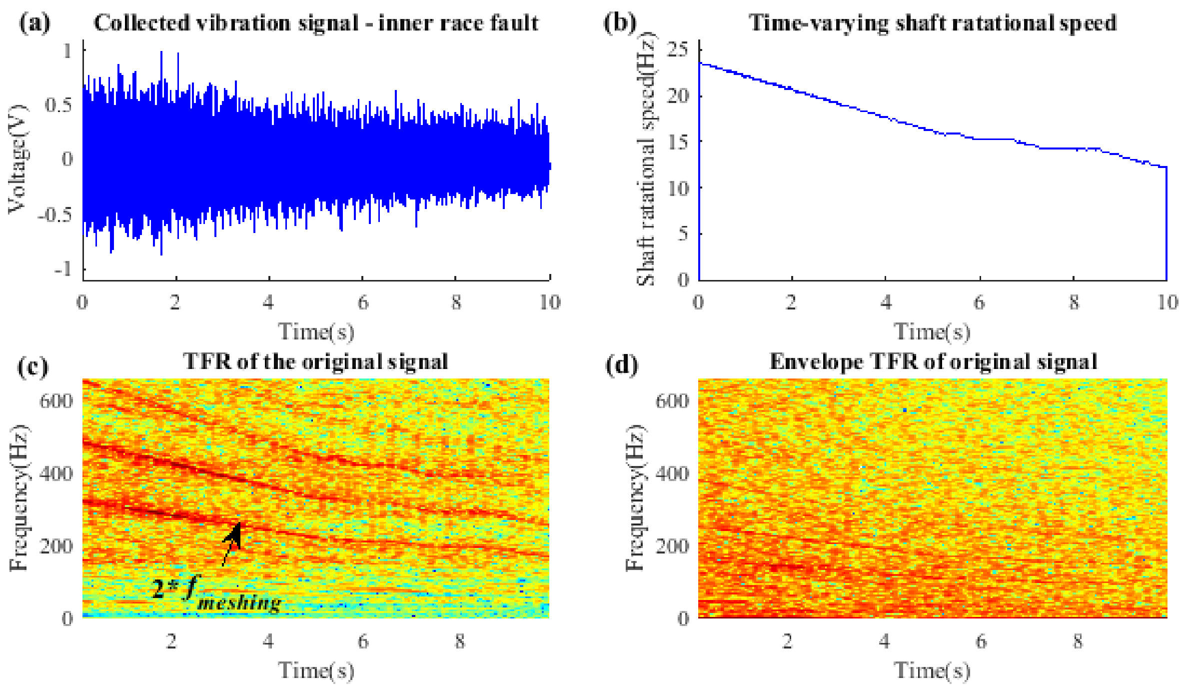

Fig. 7.14 (a) Collected vibration signal of experiment 1, (b) measured time-varying rotational speed, (c) TFR of the original signal, and (d) envelope TFR of the original signal.

Applying the fast-kurtogram to the full-length vibration signal, a kurtogram is obtained, shown in Fig. 7.15(a). According to the kurtogram, the maximum SK is located at level 2.5 (bandwidth 1667Hz) with center frequency 7500Hz. The signal is then filtered with the optimal filter and the envelope TFR of the filtered signal is obtained, shown in Fig. 7.15(b). Unlike the results obtained in the simulation and the experiment with the bearing outer race fault, in the envelope TFR obtained by the fast-kurtogram, the IFCF of the inner race fault and its harmonic can be faintly observed. The result indicates that the bearing fault features are partially extracted from the collected signal by the optimal filter. A more accurate filter should be used to fully extract the bearing fault features.

The TKR of the collected signal is obtained via the short-time kurtogram, shown in Fig. 7.15(c). The parameters of the short-time kurtogram are selected to be the same as the case in experiment 1. The window length is set at 20000 samples (1s) and the hop size is set at 12857 samples (about 0.64s). The total number of windows is 15. The parameters of the 15 obtained optimal filters, i.e. center frequency f_c and cut-off frequencies $f_c \pm \Delta f$, are shown in Fig.

7.15(d). The corresponding SK values of the optimal filters are shown in Fig. 7.15(e). According to the SK values, the last optimal filter has the maximum SK. Therefore, the last optimal filter is selected as the “final” optimal filter. The center frequency of the final optimal filter is 6250Hz with a bandwidth of 832Hz. The filter is then applied to the collected signal and the envelope TFR of the filtered signal is obtained, shown in Fig. 7.15(f). In the envelope TFR shown in Fig. 7.15(f), the IFCF of the inner race fault and its harmonics can be clearly observed. Additionally, the gear meshing frequency is fully removed from the envelope TFR. The result shows that the bearing fault features are effectively extracted from the original signal via the short-time kurtogram.

Comparing the result of the fast-kurtogram shown in Fig. 7.15(b) to the result of the short-time kurtogram shown in Fig. 7.15(f), both envelope TFR show the IFCF and its harmonics. However, in the envelope TFR obtained via the short-time kurtogram, the IFCF and its harmonics can be observed more clearly, especially in the time span [6s, 10s]. This implies that the final optimal filter obtained by the short-time kurtogram may be more effective for the extraction of the bearing fault features. The comparison reveals that the short-time kurtogram performs better than the single fast-kurtogram for the bearing fault feature extraction under time-varying rotational speed conditions.

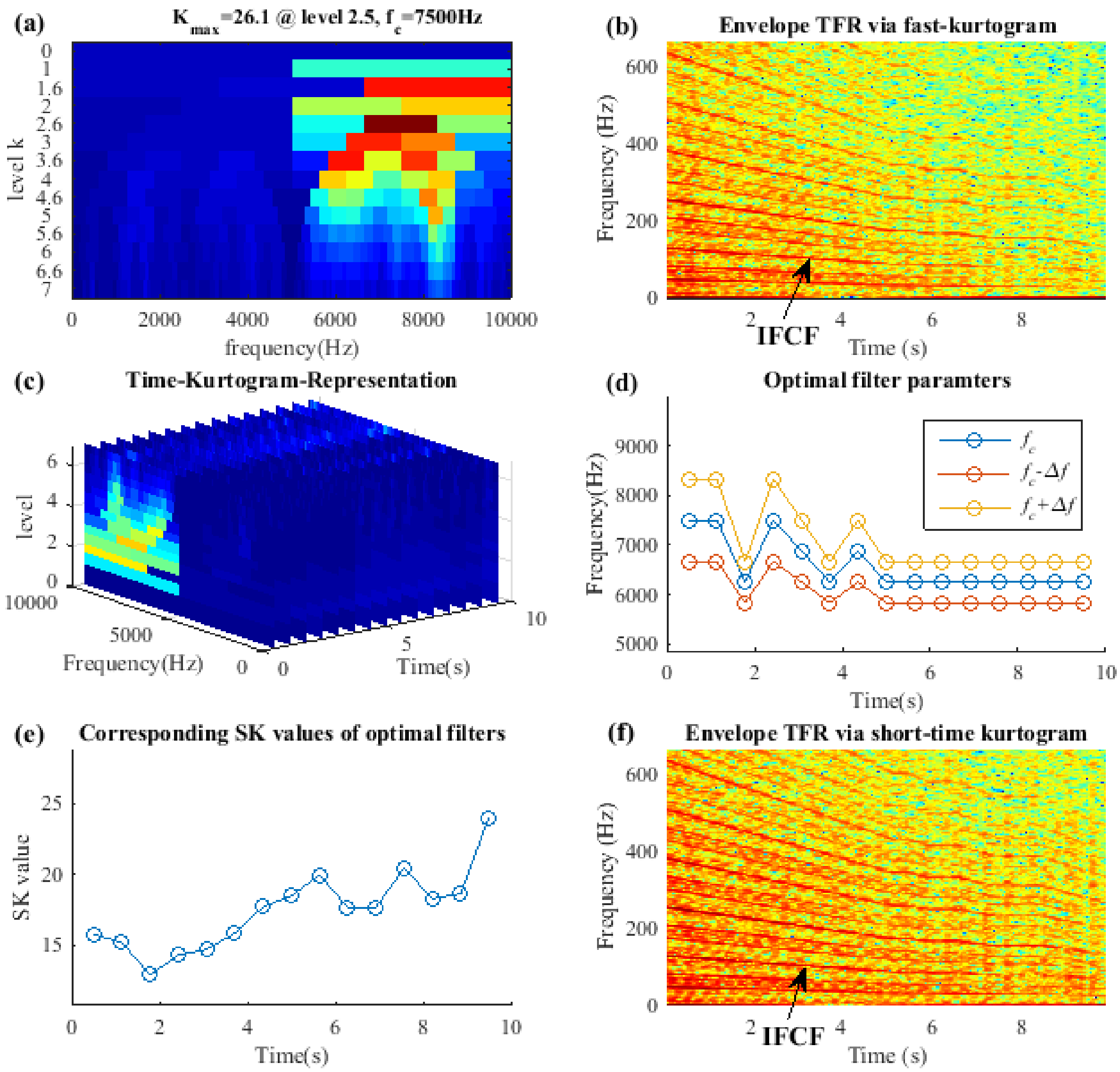

Fig. 7.15 (a) Kurtogram of the original signal obtained via the fast-kurtogram, (b) envelope TFR of the filtered signal via the fast-kurtogram, (c) TKR obtained via the short-time kurtogram, (d) parameters of a series optimal filters, € corresponding SK values of the optimal filters, and (f) envelope TFR of the filtered signal via the short-time kurtogram.

7.7 Discussion

Both in simulations and in experiments, the results show that the performance of the short-time kurtogram is better than that of the single fast-kurtogram under time-varying speed conditions. The reason is that the time-varying speed changes non-stationarities of the signal. As previously mentioned, the kurtogram locates the optimal filter by using the SK to indicate the presence of non-stationary transients within a certain band, since the bearing fault-induced signal contains non-stationary transients. Normally, the collected signal also contains interference signals, which is considered stationary under constant speed condition. However, under time-varying speed conditions, the interference signals also become non-stationary. Therefore, the effectiveness of the fast-kurtogram is suppressed. By segmenting the signal into segments of (almost) constant speed, the short-time kurtogram improves the effectiveness of the fast-kurtogram for time-varying speed cases.

To interpret the changes to a signal after segmentation, a comparison diagram is shown in Fig. 7.16. Fig. 7.16(a) is the frequency spectrum of the full-length signal used in simulation case 2 and Fig. 7.16(b) is the frequency spectrum of the first segment in simulation case 2. The frequency spectrum is obtained via Fourier transform. The rectangles in the figure represent the pass-band of optimal filters located by the fast-kurtogram. In the simulation, the expected center frequency of the optimal filter is the resonance frequency (4000Hz). In Fig. 7.16(a), due to their time-varying nature, the interference frequency (IIF) and its second harmonic (2*IIF) are distributed within a range which looks similar to the distribution of the frequencies around the resonance frequency (4000Hz). Due to the non-stationary interference signal, the center frequency of the optimal filter is located at approximately the maximum value of 2*IIF. In Fig. 7.16(b), the IIF and its second harmonic are not distributed (only single peaks) in the frequency spectrum, which is similar to the constant speed case. The center frequency of the optimal filter is approximately 4000Hz since the interference signal can be considered stationary. The comparison indicates that segmenting the signal into constant speed-like segments is the reason that the short-time kurtogram is effective under time-varying speed conditions.

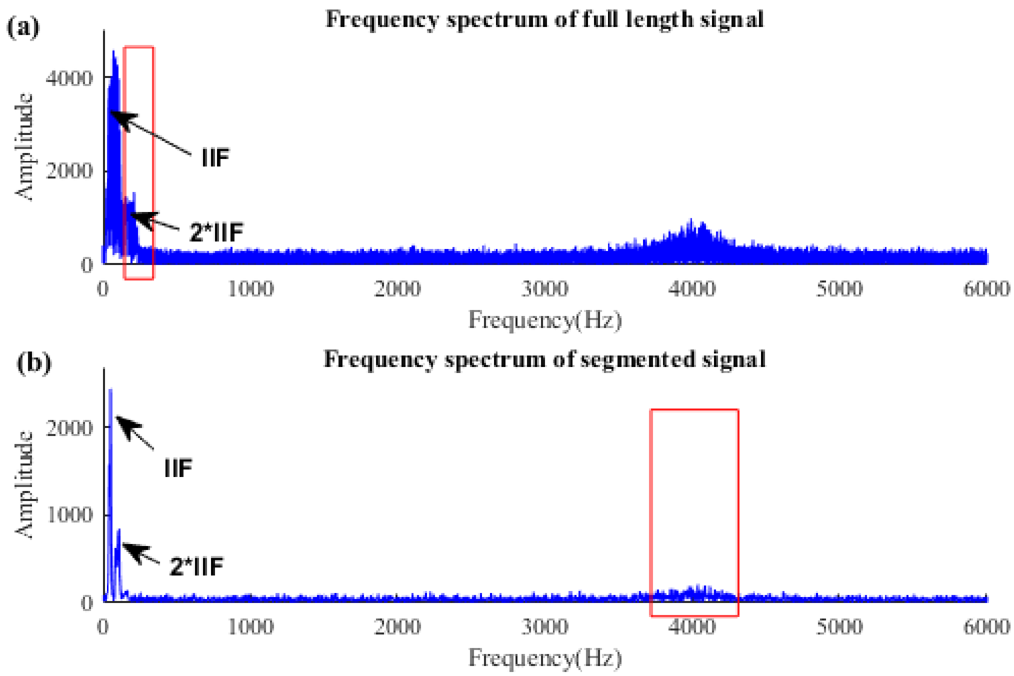

Fig. 7.16 (a) Frequency spectrum of full-length signal used in simulation case 2 and (b) frequency spectrum of the first segment in simulation case 2. The rectangle in each frequency spectrum represents the pass-band of the optimal filter located by fast-kurtogram.

It can be also seen in simulations and experiments that the results of the TKR and the corresponding optimal filters are dependent on the segment size (window size). To compare the results of the optimal filters with different window sizes, two sets of optimal filters are obtained by using the simulated signal in simulation case 2, shown in Fig. 7.17. In Fig. 7.17(a), the window size is set as 5500 (0.275 second). To keep the same amount of segments (M=20), the hop size is changed to 4973. In Fig. 7.17(b), the window size is set as 20000 (1 second) and the hop size is changed to 4210 to obtain 20 segments. The rectangles in Fig. 7.17 label the final optimal filters determined by the maximum SK value. Comparing the results shown in Fig. 7.10(d) with window size 10000 (0.5 second), more improper filters are obtained in Fig. 7.17(a) with a smaller window and more proper filters are obtained in Fig. 7.17(b) with a larger window. However, if the window is too large, such as the full-length signal, the result can be improper. A smaller window segment gives the signal more constant speed-like segments, however, the frequency resolution is reduced which may affect the accuracy of the results. On the other hand, a larger window increases the frequency resonance of the kurtogram, however, the non-stationary interference signal may affect the accuracy of the result. Fortunately, by choosing the final optimal filter as the one with the maximum SK value, the result of the short-time kurtogram can be still effective even if some improper optimal filters are obtained by the TKR.

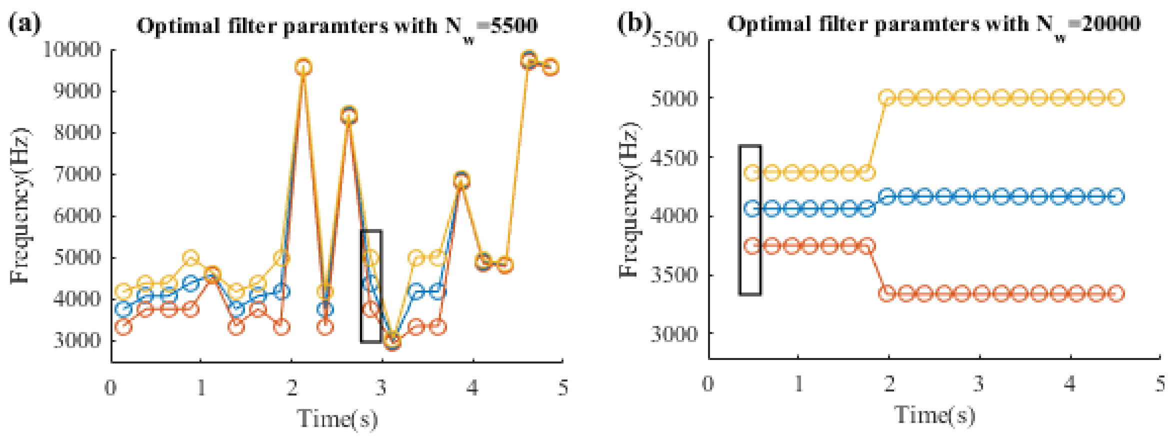

Fig. 7.17 (a) Optimal filter parameters with window size=5500 (0.275s) and (b) optimal filter parameters with window size=20000 (1s). The rectangles represent the chose final optimal filter.

7.8 Conclusions

In this chapter, a short-time kurtogram was proposed to extract bearing fault features under time-varying rotational speed conditions. The short-time kurtogram was developed by windowing the signal for a short time duration and applying the fast-kurtogram to each segment. The performance of the short-time kurtogram was tested with simulated signals. The results showed the single fast-kurtogram was unable to locate the optimal filter properly under time-varying speed conditions. However, the short-time kurtogram can be effectively used to extract bearing fault features under time-varying speed conditions. Additionally, the performance of the short-time kurtogram was examined with experimental data which were collected under time-varying speed conditions from a bearing with an outer race fault and a bearing with an inner race fault, respectively. For the experimental data, the single fast-kurtogram was not able to properly locate the optimal filter for the bearing fault feature extraction. However, the short-time kurtogram effectively extracted bearing fault features under time-varying speed conditions.

Chapter 8 Contributions, conclusions and future work

8.1 Contributions

The goal of this thesis was to address several problems associated with bearing fault diagnosis under constant speed and under time-varying speed conditions. Under constant speed conditions, the bearing faults can be diagnosed in the frequency domain by observing the FCF and its harmonics. However, the bearing fault signature needs to be extracted first, since the collected bearing vibration signal is often contaminated by random noise and interferences. Existing methods for bearing fault signature extraction is either dependent on the frequency features or relying on the acquisition of reference signals. Their effectiveness is suppressed when the frequencies of interest and the frequency of the interference are similar or the reference signal is not measurable. Thus, more reliable methods for bearing fault signature extraction need to be proposed. Under time-varying speed conditions, reliable methods for bearing fault signature extraction should be also investigated. Additionally, bearing faults cannot be directly diagnosed in the frequency domain under time-varying speed conditions, since the FCF is also time-varying. This can be solved by resampling the signal according to the speed so that the time-varying FCF becomes constant. However, the accuracy of the resampling is limited and extra instruments are needed to measure the time-varying speed. Therefore, resampling-free and tachometer-free methods are also required for bearing fault diagnosis under time-varying speed conditions.

To address the problems mentioned above, methods were investigated and new methods were proposed. The contributions of this thesis are listed as follows:

(1) The effects of parameter selection on the performance of the OBSD method for bearing fault signature extraction under constant speed conditions were investigated in Chapter 3. The OBSD method is frequency-independent and reference-free.

(2) A new method called “Auto-OBSD” was proposed in Chapter 4 to automatically select the OBSD method-related parameters for bearing fault signature extraction under constant speed conditions.

(3) A new method (MTFCE) was proposed in Chapter 5 for bearing fault diagnosis under time-varying speed conditions without the presence of interference. The proposed method is resampling-free and tachometer-free.

(4) A new resampling-free and tachometer-free method based on the OBSD and the MTFCE was proposed in Chapter 6 for bearing fault diagnosis under time-varying speed conditions with the presence of interference.

(5) A new method called short-time kurtogram was proposed in Chapter 7 for bearing fault signature extraction under time-varying speed conditions.

In all cases, the proposed methods were validated with both simulated and experimentally acquired signals.

8.2 Conclusions

8.2.1 Constant speed case

The work accomplished for bearing fault signature extraction under constant speed conditions is summarized as follows.

(1) *Effects of Parameter Selection on OBSD for Bearing Fault Signature Extraction*

The OBSD can be very effective for bearing fault signature extraction in the presence of interference signals. The effects of the selection of the OBSD parameters on the performance of the method were analyzed in Chapter 3. Additionally, the results of the implementation of the OBSD on experimental signals validated the effectiveness of the analysis. Parameter selection for the application of the OBSD to bearing fault signature extraction can now be carried out based on the discussed effects.

(2) *Auto-OBSD with an application to bearing fault signature extraction*

An automatic parameter selection method was developed for the OBSD for bearing fault signature extraction in Chapter 4 with three index guided self-tuning algorithms. A convergence index was proposed to guide the self-tuning of two parameters. Additionally, the Smoothness Index (SI) was employed to guide the automatic selection of two other parameters. By applying the proposed auto-OBSD to a series of simulated signals, it was found that the method performed well even when the interference frequency is almost identical to the fault characteristic frequency and when the signal-to-interference ratio is very low. Furthermore, the method was implemented on experimental data collected from a bearing with outer race fault and a bearing with inner race fault, respectively. The results indicate that the method is able to extract the bearing fault signature from the signal contaminated by gear meshing signals for both the outer race case and inner race case. The above results were achieved without any pre-denoising steps.

The proposed auto-OBSD method is robust to the variation of the rotational speed. The reason for this is that the auto-OBSD is proposed based on the OBSD method, which decomposes a signal according to oscillatory behavior rather than frequency band. The result is independent on the rotational speed. Additionally, the effectiveness of the OBSD method for bearing fault signature extraction under time-varying speed conditions has been validated in chapter 6. Therefore, the auto-OBSD is potentially effective under time-varying speed conditions.

8.2.2 Time-varying speed case

The research completed for bearing fault diagnosis under time-varying speed conditions is summaried in the following.

(1) *bearing fault diagnosis under unknown time-varying rotational speed conditions via multiple time-frequency curve extraction (MTFCE)*

A multiple T-F curve extraction algorithm was proposed in Chapter 5 to extract T-F curves from the TFR for bearing fault diagnosis under time-varying speed conditions. Based on the multiple T-F curve extraction algorithm, a procedure for bearing fault diagnosis under time-varying speed conditions was proposed which is resampling-free and tachometer-free. The performance of the proposed procedure was tested with four types of simulated signals. The results of simulations illustrate that the proposed procedure is effective. The effectiveness of the proposed procedure was also validated with experimental data. The experimental data were collected from a bearing with an outer race fault, a bearing with an inner race fault and a healthy bearing, respectively. The results reveal that the proposed signal can be effectively used for bearing fault diagnosis under time-varying speed conditions.

(2) *Tachometer-free and resampling-free bearing fault diagnosis under time-varying speed conditions*

A tachometer-free and resampling-free method was proposed for bearing fault diagnosis under time-varying speed conditions in the presence of interference signals in Chapter 6. To properly extract the true ISRF, an ISRF-search diagnosis strategy was proposed for this method, as a crucial component of the proposed joint OBSD and MTFCE algorithm. The performance of the proposed method was examined using four different types of simulated signals. The results show that the proposed method can be used for effective bearing fault diagnosis under time-varying speed conditions in the presence of interference. The effectiveness of the proposed method was also validated by signals collected from a test rig operated under time-varying speed conditions. The collected data contains not only the bearing vibrations signal but also the gear meshing signal, which is considered as an interference signal for bearing fault diagnosis. Signals were collected from two experiments conducted under different conditions. It was demonstrated that the fault can be effectively diagnosed in both experiments using the proposed method.

(3) *short-time kurtogram for bearing fault feature extraction under time-varying speed conditions*

A short-time kurtogram was proposed to extract bearing fault features under time-varying rotational speed conditions in Chapter 7. The short-time kurtogram was developed by windowing the signal for a short time duration and applying the fast-kurtogram to each segment. The performance of the short-time kurtogram was tested with simulated signals. The results showed that the short-time kurtogram can be effectively used to extract bearing fault features under time-varying speed conditions. Additionally, the performance of the short-time kurtogram was examined with experimental data which were collected under time-varying speed conditions from a bearing with an outer race fault and a bearing with an inner race fault, respectively. From the results, the short-time kurtogram effectively extracted bearing fault features under time-varying speed conditions

A comparison of the methods proposed for the case of time-varying speed is summarized in Table 8.1, including their advantages and disadvantages. They can be selected for applications according to the characteristics listed in the table.

Table 8.1 Comparison of methods proposed for time-varying speed conditions

Method	Pros	Cons
Proposed in chapter 5 (MTFCE)	Can be used to diagnose bearing faults directly	May not be effective if the signal is smeared by interference signals
Proposed in chapter 6 (OBSD+MTFCE)	Can be used to diagnose bearing faults directly, effective with the presence of interference signals	Relatively high computational cost
Proposed in chapter 7 (Short-time kurtogram)	Can be used to extract bearing fault features with the presence of interference signals, Low computational cost	Cannot be used to diagnose bearing faults directly

8.3 Future work

Potential future work is listed as follows.

(1) *Extension of the auto-OBSD for bearing fault signature extraction under time-varying speed conditions*

Theoretically, the auto-OBSD can be used for fault signature extraction under time-varying speed conditions, since the OBSD method is frequency-independent. The performance of the auto-OBSD for the time-varying speed cases should be tested. New methods may be proposed by modifying the auto-OBSD according to the performance.

(2) *Extension of the multiple time-frequency curve extraction (MTFCE) to other applications*

The MTFCE algorithm can be used to extract multiple time-frequency curves from the TFR of a signal. The signal is not limited to bearing vibrations signal. The method is potentially promising for other applications, such as fault diagnosis for gearboxes under time-varying speed conditions.

(3) *Extension of the short-time kurtogram to other applications*

Similarly, the application of the short-time kurtogram is not limited to bearing fault feature extraction under time-varying speed condition. It can also be used to filter other signals other than the bearing vibration signal collected under time-varying speed conditions.

References

[1] A. Heng, S. Zhang, A. C. C. Tan, and J. Mathew, "Rotating machinery prognostics: State of the art, challenges and opportunities," *Mech. Syst. Signal Process.*, vol. 23, no. 3, pp. 724–739, 2009.

[2] F. K. Geitner and H. P. Bloch, *Practical Machinery Management for Process Plants: Volume 2: Machinery Failure Analysis and Troubleshooting*. Gulf Professional Publishing, 1997.

[3] G. K. Singh and S. A. S. Al Kazzaz, "Induction machine drive condition monitoring and diagnostic research- a survey," *Electr. Power Syst. Res.*, vol. 64, no. 2, pp. 145–158, 2003.

[4] M. R. Hoeprich, "Rolling-Element Bearing Internal Temperatures," *Tribol. Trans.*, vol. 39, no. 4, pp. 855–858, 1996.

[5] R. Schulz, S. Verstockt, J. Vermeiren, M. Loccufier, K. Stockman, and S. Van Hoecke, "Thermal Imaging for Monitoring Rolling Element Bearings," in *12th International conference on Quantitative InfraRed Thermography*, 2014.

[6] W. M. Hannon, "Rolling-Element Bearing Heat Transfer—Part I: Analytic Model," *ASME J. Tribol.*, vol. 137, no. 3, pp. 311021–3110211, 2015.

[7] P. D. McFadden and J. D. Smith, "Vibration monitoring of rolling element bearings by the high-frequency resonance technique - a review," *Tribol. Int.*, vol. 17, pp. 3–10, 1984.

[8] I. Antoniadis and G. Glossiotis, "Cyclostationary Analysis of Rolling-Element Bearing Vibration Signals," *J. Sound Vib.*, vol. 248, no. 5, pp. 829–845, 2001.

[9] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics-A tutorial," *Mech. Syst. Signal Process.*, vol. 25, no. 2, pp. 485–520, 2011.

[10] D. V. Spyropoulos and E. D. Mitronikas, "A review on the faults of electric machines used in electric ships," *Adv. Power Electron.*, vol. 2013, 2013.

[11] A. Choudhury and N. Tandon, "Application of acoustic emission technique for the detection of defects in rolling element bearings," *Tribol. Int.*, vol. 33, no. 1, pp. 39–45,

2000.

[12] X. Liu, X. Wu, and C. Liu, “A comparison of acoustic emission and vibration on bearing fault detection,” *2011 Int. Conf. Transp. Mech. Electr. Eng.*, pp. 922–926, 2011.

[13] F. Hemmati, M. Alqaradawi, and M. S. Gadala, “Rolling element bearing fault diagnostics using acoustic emission technique and advanced signal processing,” *Proc. Inst. Mech. Eng. Part J J. Eng. Tribol.*, vol. 0, no. 0, pp. 1–14, 2015.

[14] J. L. Miller and D. Kitaljevich, “In-line oil debris monitor for aircraft engine condition assessment,” *2000 IEEE Aerosp. Conf. Proc. (Cat. No.00TH8484)*, vol. 6, no. 860, pp. 49–56, 2000.

[15] C. Radu, “The Most Common Causes of Bearing Failure and the Importance of Bearing Lubrication,” *RKB Tech. Rev.*, vol. February, 2010.

[16] C. Li and M. Liang, “Extraction of oil debris signature using integral enhanced empirical mode decomposition and correlated reconstruction,” *Meas. Sci. Technol.*, vol. 22, no. 8, p. 85701, 2011.

[17] N. Tandon and A. Choudhury, “Review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings,” *Tribol. Int.*, vol. 32, no. 8, pp. 469–480, 1999.

[18] N. G. Nikolaou and I. A. Antoniadis, “Demodulation of Vibration Signals Generated By Defects in Rolling Element Bearings Using Complex Shifted Morlet Wavelets,” *Mech. Syst. Signal Process.*, vol. 16, no. 4, pp. 677–694, 2002.

[19] D. Dyer and R. M. Stewart, “Detection of Rolling Element Bearing Damage by Statistical Vibration Analysis,” vol. 100, no. 2, pp. 229–235, 1978.

[20] R. B. W. Heng and M. J. M. Nor, “Statistical analysis of sound and vibration signals for monitoring rolling element bearing condition,” *Appl. Acoust.*, vol. 53, no. 1–3, pp. 211–226, 1998.

[21] S. A. Aye, “Statistical Approach for Tapered Bearing Fault Detection using Different Methods,” vol. III, pp. 6–9, 2011.

[22] B. Samanta, K. R. Al-Balushi, and S. A. Al-Araimi, “Artificial neural networks and

support vector machines with genetic algorithm for bearing fault detection," *Eng. Appl. Artif. Intell.*, vol. 16, no. 7–8, pp. 657–665, 2003.

[23] J. Antoni, "The spectral kurtosis: A useful tool for characterising non-stationary signals," *Mech. Syst. Signal Process.*, vol. 20, pp. 282–307, 2006.

[24] J. Antoni and R. B. Randall, "The spectral kurtosis: Application to the vibratory surveillance and diagnostics of rotating machines," *Mech. Syst. Signal Process.*, vol. 20, pp. 308–331, 2006.

[25] I. S. Bozchalooi and M. Liang, "A smoothness index-guided approach to wavelet parameter selection in signal de-noising and fault detection," *J. Sound Vib.*, vol. 308, pp. 246–267, 2007.

[26] I. S. Bozchalooi and M. Liang, "A joint resonance frequency estimation and in-band noise reduction method for enhancing the detectability of bearing fault signals," *Mech. Syst. Signal Process.*, vol. 22, pp. 915–933, 2008.

[27] G. K. Chaturvedi and D. W. Thomas, "Bearing Fault Detection Using Adaptive Noise Cancelling," *J. Mech. Des.*, vol. 104, no. April 1982, p. 280, 1982.

[28] I. Khemili and M. Chouchane, "Detection of rolling element bearing defects by adaptive filtering," *Eur. J. Mech. A/Solids*, vol. 24, no. October 2004, pp. 293–303, 2005.

[29] T. Wang, M. Liang, J. Li, W. Cheng, and C. Li, "Bearing fault diagnosis under unknown variable speed via gear noise cancellation and rotational order sideband identification," *Mech. Syst. Signal Process.*, vol. 62, pp. 30–53, 2015.

[30] M. Bai, J. Huang, M. Hong, and F. Su, "Fault diagnosis of rotating machinery using an intelligent order tracking system," *J. Sound Vib.*, vol. 280, pp. 699–718, 2005.

[31] P. Borghesani, P. Pennacchi, R. B. Randall, and R. Ricci, "Order tracking for discrete-random separation in variable speed conditions," *Mech. Syst. Signal Process.*, vol. 30, pp. 1–22, 2012.

[32] T. Wang, M. Liang, J. Li, and W. Cheng, "Rolling element bearing fault diagnosis via fault characteristic order (FCO) analysis," *Mech. Syst. Signal Process.*, vol. 45, no. 1, pp. 139–153, 2014.

[33] K. M. Bossley, R. J. McKendrick, C. J. Harris, and C. Mercer, “Hybrid computed order tracking,” *Mech. Syst. Signal Process.*, vol. 13, no. 4, pp. 627–641, 1999.

[34] K. R. Fyfe and E. D. S. Munck, “Analysis of computed order tracking,” *Mech. Syst. Signal Process.*, vol. 11, no. 2, pp. 187–205, 1997.

[35] I. S. Bozchalooi and M. Liang, “Identification of the high SNR frequency band for bearing fault signature enhancement,” *Proc. 14th Int. Conf. Mechatronics Mach. Vis. Pract. M2VIP2007*, pp. 39–43, 2007.

[36] M. Liang and I. Soltani Bozchalooi, “An energy operator approach to joint application of amplitude and frequency-demodulations for bearing fault detection,” *Mech. Syst. Signal Process.*, vol. 24, no. 5, pp. 1473–1494, 2010.

[37] A. Djebala, N. Ouelaa, and N. Hamzaoui, “Detection of rolling bearing defects using discrete wavelet analysis,” *Meccanica*, vol. 43, no. 3, pp. 339–348, 2008.

[38] T. G. Habetler and R. G. Harley, “Incipient Bearing Fault Detection via Motor Stator Current Noise Cancellation Using Wiener Filter,” *IEEE Trans. Ind. Appl.*, vol. 45, no. 4, pp. 1309–1317, 2009.

[39] J. Antoni and R. B. Randall, “Unsupervised noise cancellation for vibration signals: Part I - Evaluation of adaptive algorithms,” *Mech. Syst. Signal Process.*, vol. 18, pp. 89–101, 2004.

[40] I. W. Selesnick, “Resonance-based signal decomposition: A new sparsity-enabled signal analysis method,” *Signal Processing*, vol. 91, no. 12, pp. 2793–2809, 2011.

[41] S. Hou, M. Liang, Y. Zhang, and C. Li, “Vibration signal demodulation and bearing fault detection: A clustering-based segmentation method,” *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.*, vol. 0, no. 0, pp. 1–12, 2013.

[42] D. Benitez, P. A. Gaydecki, A. Zaidi, and A. P. Fitzpatrick, “The use of the Hilbert transform in ECG signal analysis,” *Comput. Biol. Med.*, vol. 31, pp. 399–406, 2001.

[43] T. Barszcz and A. Jabłoński, “A novel method for the optimal band selection for vibration signal demodulation and comparison with the Kurtogram,” *Mech. Syst. Signal Process.*, vol. 25, no. 1, pp. 431–451, 2011.

[44] H. Hong and M. Liang, “K-Hybrid: A Kurtosis-Based Hybrid Thresholding Method for Mechanical Signal Denoising,” *J. Vib. Acoust.*, vol. 129, no. August, pp. 458–470, 2007.

[45] Y. Zhang, M. Liang, C. Li, and S. Hou, “A Joint Kurtosis-Based Adaptive Bandstop Filtering and Iterative Autocorrelation Approach to Bearing Fault Detection,” *J. Vib. Acoust.*, vol. 135, no. October 2013, p. 51026, 2013.

[46] R. Dwyer, “Detection of non-Gaussian signals by frequency domain Kurtosis estimation,” *ICASSP ’83. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 8, pp. 607–610, 1983.

[47] P. K. Kankar, S. C. Sharma, and S. P. Harsha, “Rolling element bearing fault diagnosis using wavelet transform,” *Neurocomputing*, vol. 74, no. 10, pp. 1638–1645, 2011.

[48] A. Klepka, “Wavelet based signal demodulation technique for bearing fault detection,” *Mech. Mech. Eng.*, vol. 15, no. 4, pp. 63–71, 2011.

[49] C. K. E. Nizwan, S. A. Ong, M. F. M. Yusof, and M. Z. Baharom, “A wavelet decomposition analysis of vibration signal for bearing fault detection,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 50, p. 12026, 2013.

[50] Y. Wang and M. Liang, “Identification of multiple transient faults based on the adaptive spectral kurtosis method,” *J. Sound Vib.*, vol. 331, no. 2, pp. 470–486, 2012.

[51] H. Faghidi, “Non-parametric and Non-filtering Methods for Rolling Element Bearing Condition Monitoring,” University of Ottawa, 2014.

[52] H. Faghidi and M. Liang, “Detection of Bearing Fault Detection From Heavily Contaminated Signals: A Higher-Order Analytic Energy Operator Method,” *J. Vib. Acoust.*, vol. 137, no. 4, p. 41012, 2015.

[53] M. Liang and H. Faghidi, “Intelligent bearing fault detection by enhanced energy operator,” *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7223–7234, 2014.

[54] M. Liang and H. Faghidi, “A non-parametric non-filtering approach to bearing fault detection in the presence of multiple interference,” *Meas. Sci. Technol.*, vol. 24, p. 105013, 2013.

[55] S. Haykin, *Adaptive Filter Theory*, 5th ed. Pearson Education, 2015.

[56] R. A. David, S. D. Stearns, and G. R. Elliott, “IIR algorithms for adaptive line

enhancement," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP*, 1983, pp. 17–21.

[57] J. Antoni and R. B. Randall, "Unsupervised noise cancellation for vibration signals: Part I - Evaluation of adaptive algorithms," *Mech. Syst. Signal Process.*, vol. 18, no. 1, pp. 89–101, 2004.

[58] J. Shiroishi, Y. Li, S. Liang, T. Kurfess, and S. Danyluk, "Bearing condition diagnostics via vibration and acoustic emission measurements," *Mech. Syst. Signal Process.*, vol. 11, no. 5, pp. 693–705, 1997.

[59] J. Antoni and R. B. Randall, "Unsupervised noise cancellation for vibration signals: Part II - A novel frequency-domain algorithm," *Mech. Syst. Signal Process.*, vol. 18, pp. 103–117, 2004.

[60] E. Parzen, "Autoregressive Spectral Estimation," *Handb. Stat.*, vol. 3, pp. 221–247, 1983.

[61] J. Altmann and J. Mathew, "Multiple Band-Pass Autoregressive Demodulation for Rolling-Element Bearing Fault Diagnosis," *Mech. Syst. Signal Process.*, vol. 15, no. 5, pp. 963–977, 2001.

[62] I. W. Selesnick, "Wavelet transform with tunable Q-factor," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3560–3575, 2011.

[63] I. W. Selesnick, "Sparse signal representations using the tunable Q-factor wavelet transform," *Proc. SPIE Wavelets Sparsity XIV*, vol. 8131, pp. 1–15, 2011.

[64] J. Luo, D. Yu, and M. Liang, "A kurtosis-guided adaptive demodulation technique for bearing fault detection based on tunable-Q wavelet transform," *Meas. Sci. Technol.*, vol. 24, p. 55009, 2013.

[65] W. He, Y. Zi, B. Chen, S. Wang, and Z. He, "Tunable Q-factor wavelet transform denoising with neighboring coefficients and its application to rotating machinery fault diagnosis," *Sci. China Technol. Sci.*, vol. 56, no. 8, pp. 1956–1965, 2013.

[66] W. He, Y. Zi, B. Chen, F. Wu, and Z. He, "Automatic fault feature extraction of mechanical anomaly on induction motor bearing using ensemble super-wavelet transform," *Mech. Syst. Signal Process.*, vol. 54–55, pp. 457–480, 2015.

[67] L. Cui, D. Mo, H. Wang, and P. Chen, "Resonance-Based Nonlinear Demodulation Analysis Method of Rolling Bearing Fault," *Adv. Mech. Eng.*, vol. 5, pp. 1–13, 2013.

[68] H. Wang, J. Chen, and G. Dong, "Feature extraction of rolling bearing's early weak fault based on EEMD and tunable Q-factor wavelet transform," *Mech. Syst. Signal Process.*, vol. 48, no. 1–2, pp. 103–119, 2014.

[69] J. Shi and M. Liang, "Intelligent bearing fault signature extraction via iterative oscillatory behavior based signal decomposition (IOBSD)," *Expert Syst. Appl.*, vol. 45, pp. 40–55, 2016.

[70] Z. Feng, M. Liang, and F. Chu, "Recent advances in time-frequency analysis methods for machinery fault diagnosis: A review with application examples," *Mech. Syst. Signal Process.*, vol. 38, no. 1, pp. 165–205, 2013.

[71] P. Schniter, "Short-time Fourier Transform," *Transform*, vol. 2, pp. 2–5, 2003.

[72] Q. Sun and Y. Tang, "Singularity Analysis Using Continuous Wavelet Transform for Bearing Fault Diagnosis," *Mech. Syst. Signal Process.*, vol. 16, pp. 1025–1041, 2002.

[73] P. Borghesani, R. Ricci, S. Chatterton, and P. Pennacchi, "A new procedure for using envelope analysis for rolling element bearing diagnostics in variable operating conditions," *Mech. Syst. Signal Process.*, vol. 38, no. 1, pp. 23–35, 2013.

[74] M. Cocconcelli, R. Zimroz, R. Rubini, and W. Bartelmus, "STFT based approach for ball bearing fault detection in a varying speed motor," *Cond. Monit. Mach. Non-Stationary Oper.*, pp. 41–50, 2012.

[75] N. Baydar and A. Ball, "A comparative study of acoustic and vibration signals in detection of gear failures using Wigner-Ville distribution," *Mech. Syst. Signal Process.*, vol. 15, no. 6, pp. 1091–1107, 2001.

[76] S. Rajagopalan, J. a. Restrepo, J. M. Aller, T. G. Habetler, and R. G. Harley, "Nonstationary motor fault detection using recent quadratic time-frequency representations," *IEEE Trans. Ind. Appl.*, vol. 44, no. 3, pp. 735–744, 2008.

[77] P. Flandrin, L. C. France, O. Rioul, C. P. Abbpe, and G. Leclerc, "Affine smoothing of the Wigner-Ville distribution," in *1990 International Conference onAcoustics, Speech,*

and Signal Processing, 1990. ICASSP-90., 1990, no. 1, pp. 2455–2458.

[78] I. Djurović and L. Stanković, “Time-frequency representation based on the reassigned S-method,” *Signal Processing*, vol. 77, no. 1, pp. 115–120, 1999.

[79] D. Gabor, “Theory of communication,” *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, vol. 93, no. 26. pp. 429–441, 1946.

[80] J. Shi, M. Liang, D.-S. Necsulescu, and Y. Guan, “Generalized stepwise demodulation transform and synchrosqueezing for time–frequency analysis and bearing fault diagnosis,” *J. Sound Vib.*, vol. 368, pp. 202–222, 2016.

[81] J. Shi, M. Liang, and Y. Guan, “Bearing fault diagnosis under variable rotational speed via the joint application of windowed fractal dimension transform and generalized demodulation: A method free from prefiltering and resampling,” *Mech. Syst. Signal Process.*, vol. 68–69, pp. 15–33, 2016.

[82] N. Hess-Nielsen and M. V. Wickerhauser, “Wavelets and time-frequency analysis,” *Proc. IEEE*, vol. 84, no. 4, pp. 523–540, 1996.

[83] R. Zhou, W. Bao, N. Li, X. Huang, and D. R. Yu, “Mechanical equipment fault diagnosis based on redundant second generation wavelet packet transform,” *Digit. Signal Process. A Rev. J.*, vol. 20, no. 1, pp. 276–288, 2010.

[84] Y. Wang and M. Liang, “An adaptive SK technique and its application for fault detection of rolling element bearings,” *Mech. Syst. Signal Process.*, vol. 25, no. 5, pp. 1750–1764, 2011.

[85] J. Shi and M. Liang, “Oscillatory behavior based fault feature extraction for bearing fault diagnosis,” in *Proceedings of the 2015 International Conference on Advanced Mechatronic Systems*, 2015, pp. 473–478.

[86] M. V Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, “Fast image recovery using variable splitting and constrained optimization,” *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2345–56, 2010.

[87] Spectra Quest Tech Note, “The Effect of Frequency Resolution in Bearing Fault Studies,” 2010.

[88] S. Stanczak, M. Wiczanowski, and H. Boche, *Fundamentals of resource allocation in wireless networks: theory and algorithms*. Springer Science & Business Media, 2009.

[89] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends® Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.

[90] J. C. Brown, "An efficient algorithm for the calculation of a constant Q transform," *J. Acoust. Soc. Am.*, vol. 92, no. 5, pp. 2698–2701, 1992.

[91] J.-L. Starck, J. Bobin, M. Elad, Y. Moudden, and D. L. Donoho, "Morphological component analysis," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 2675–2681, 2007.

[92] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. Urbana : University of Illinois Press, 1949.

[93] Y. Wang, G. Xu, A. Luo, L. Liang, and K. Jiang, "An online tacholess order tracking technique based on generalized demodulation for rolling bearing fault detection," *J. Sound Vib.*, vol. 367, pp. 233–249, 2016.

[94] C. J. Stander and P. S. Heyns, "Transmission path phase compensation for gear monitoring under fluctuating load conditions," *Mech. Syst. Signal Process.*, vol. 20, no. 7, pp. 1511–1522, 2006.

[95] I. Daubechies, J. Lu, and H. T. Wu, "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 2, pp. 243–261, 2011.

[96] C. Li and M. Liang, "A generalized synchrosqueezing transform for enhancing signal time-frequency representation," *Signal Processing*, vol. 92, no. 9, pp. 2264–2274, 2012.

[97] Z. Feng, X. Chen, M. Liang, and F. Ma, "Time-frequency demodulation analysis based on iterative generalized demodulation for fault diagnosis of planetary gearbox under nonstationary conditions," *Mech. Syst. Signal Process.*, vol. 62, pp. 54–74, 2015.

[98] Z. Feng, X. Chen, and M. Liang, "Iterative generalized synchrosqueezing transform for fault diagnosis of wind turbine planetary gearbox under nonstationary conditions," *Mech. Syst. Signal Process.*, vol. 52–53, pp. 360–375, 2015.

[99] Z. Feng, X. Chen, and M. Liang, "Joint envelope and frequency order spectrum analysis based on iterative generalized demodulation for planetary gearbox fault diagnosis under nonstationary conditions," *Mech. Syst. Signal Process.*, vol. 76–77, pp. 242–264, 2015.

[100] M. Borda, I. Nafornita, D. Isar, and A. Isar, "New instantaneous frequency estimation method based on image processing techniques," *Proc. of SPIE'05*, vol. 14, no. 2, pp. 1–11, 2005.

[101] J. Terrien, C. Marque, and G. Germain, "Ridge extraction from the time-frequency representation (TFR) of signals based on an image processing approach: Application to the analysis of uterine electromyogram AR TFR," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 5, pp. 1496–1503, 2008.

[102] M. Liebling, T.-F. Bernhard, A. H. Bachmann, L. Froehly, T. Lasser, and M. Unser, "Continuous wavelet transform ridge extraction for spectral interferometry imaging," *Prog. Biomed. Opt. Imaging - Proc. SPIE*, vol. 5690, pp. 397–402, 2005.

[103] D. Iatsenko, P. V. E. Mcclintock, and A. Stefanovska, "Extraction of instantaneous frequencies from ridges in time – frequency representations of signals," *Signal Processing*, vol. 125, pp. 290–303, 2016.

[104] R. N. Bracewell, *The Fourier Transform and its Applications*. New York: McGraw-Hill, 1978.

[105] D. BRIE, "Modelling of the Spalled Rolling Element Bearing Vibration Signal: an Overview and Some New Results," *Mech. Syst. Signal Process.*, vol. 14, no. 3, pp. 353–369, 2000.

[106] H. Huang, N. Baddour, and M. Liang, "Auto-OBSD: Automatic parameter selection for reliable Oscillatory Behavior-based Signal Decomposition with an application to bearing fault signature extraction," *Mech. Syst. Signal Process.*, vol. 86, pp. 237–259, 2017.

[107] Y. Guo, T. W. Liu, J. Na, and R. F. Fung, "Envelope order tracking for fault detection in rolling element bearings," *J. Sound Vib.*, vol. 331, no. 25, pp. 5644–5654, 2012.

[108] M. Zhao, J. Lin, X. Xu, and Y. Lei, "Tacholess envelope order analysis and its

application to fault detection of rolling element bearings with varying speeds," *Sensors*, vol. 13, no. 8, pp. 10856–10875, 2013.

[109] Y. Wang, G. Xu, A. Luo, L. Liang, and K. Jiang, "An online tacholess order tracking technique based on generalized demodulation for rolling bearing fault detection," *J. Sound Vib.*, vol. 367, pp. 233–249, 2016.

[110] S. Olhede and A. T. Walden, "A generalized demodulation approach to time-frequency projections for multicomponent signals," *Proc. R. Soc. A Math. Phys. Eng. Sci.*, vol. 461, no. 2059, pp. 2159–2179, 2005.

[111] H. Huang, N. Baddour, and M. Liang, "Algorithm for multiple time-frequency curve extraction from time-frequency representation of vibration signals for bearing fault diagnosis under time-varying speed conditions," in *Proceedings of the ASME Design Engineering Technical Conference*, 2017, pp. 1–7.

[112] J. Antoni, "Fast computation of the kurtogram for the detection of transient faults," *Mech. Syst. Signal Process.*, vol. 21, pp. 108–124, 2007.

[113] W. He, Z. N. Jiang, and K. Feng, "Bearing fault detection based on optimal wavelet filter and sparse code shrinkage," *Meas. J. Int. Meas. Confed.*, vol. 42, no. 7, pp. 1092–1102, 2009.

[114] W. Su, F. Wang, H. Zhu, Z. Zhang, and Z. Guo, "Rolling element bearing faults diagnosis based on optimal Morlet wavelet filter and autocorrelation enhancement," *Mech. Syst. Signal Process.*, vol. 24, no. 5, pp. 1458–1472, 2010.

[115] P. K. Kankar, S. C. Sharma, and S. P. Harsha, "Rolling element bearing fault diagnosis using wavelet transform," *Neurocomputing*, vol. 74, no. 10, pp. 1638–1645, 2011.

[116] N. Sawalhi, R. B. Randall, and H. Endo, "The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis," *Mech. Syst. Signal Process.*, vol. 21, pp. 2616–2633, 2007.

[117] Y. Lei, J. Lin, Z. He, and Y. Zi, "Application of an improved kurtogram method for fault diagnosis of rolling element bearings," *Mech. Syst. Signal Process.*, vol. 25, no. 5, pp. 1738–1749, 2011.

[118] D. Wang, P. W. Tse, and K. L. Tsui, “An enhanced Kurtogram method for fault diagnosis of rolling element bearings,” *Mech. Syst. Signal Process.*, vol. 35, no. 1–2, pp. 176–199, 2013.

[119] A. Mertins, *Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications*. Wiley, 1999.

[120] X. Gu, S. Yang, Y. Liu, and R. Hao, “Rolling element bearing faults diagnosis based on kurtogram and frequency domain correlated kurtosis,” *Meas. Sci. Technol.*, vol. 27, no. 12, 2016.

[121] H. Packard, “The Fundamentals of Signal Analysis,” *HP Appl. note 243*, p. 68, 1985.

[122] A. D. (Ed. . Poularikas, *Transforms and applications handbook*. CRC press, 2010.

[123] I. Selesnick, “Sparse Signal Restoration,” *Connexions*, pp. 1–13, 2009.

[124] J. Sherman and W. J. Morrison, “Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix,” *Ann. Math. Stat.*, vol. 21, no. 1, pp. 124–127, 1950.

[125] M. S. Bartlett, “An Inverse Matrix Adjustment Arising in Discriminant Analysis,” *Ann. Math. Stat.*, vol. 22, pp. 107–111, 1951.

Appendix A: Tunable Q-factor wavelet transform

The Tunable Q-factor Wavelet Transform (TQWT) of a signal can be obtained by repeatedly applying the filter bank shown in Fig. 4.6.

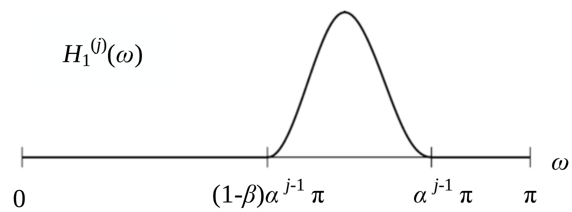

Fig. A.1 The subband-j frequency response $H_1^{(j)}(\omega)$ (cited from [62])

The frequency response of the jth sub-band generated by the filter bank is shown in Fig. A.1. It is nonzero in the interval (ω_1, ω_2) where

$$\omega_1 = (1-\beta)\alpha^{j-1}\pi, \quad \omega_2 = \alpha^{j-1}\pi \tag{A.1}$$

The center frequency and the band-width can be obtained:

$$\omega_c = \frac{1}{2}\ \omega_1 + \omega_2 \ = \alpha^j \frac{2-\beta}{2\alpha}\pi \tag{A.2}$$

$$\mathrm{BW} = \frac{1}{2}\ \omega_2 - \omega_1 \ = \frac{1}{2}\beta\alpha^{j-1}\pi \tag{A.3}$$

Then, the Q-factor is given by

$$Q = \frac{\omega_c}{BW} = \frac{2-\beta}{\beta} \tag{A.4}$$

The sampling rate of jth sub-band is given by $\beta\alpha^{j-1} f_s$, where f_s is the sampling rate of the input signal. Then, the sum of the sampling rate over all sub-bands is $\beta/(1-\alpha)\, f_s$. Hence, the oversampling rate (redundacy) is given by

$$r = \frac{\beta}{1-\alpha} \tag{A.5}$$

Therefore, with the selected parameters Q and r for TWQT, the filter bank parameters can be expressed as

$$\beta = \frac{2}{Q+1}, \quad \alpha = 1 - \frac{\beta}{r} \tag{A.6}$$

The realization of the filter bank is based on a discrete signal of finite length. An example of the behavior of the two-channel analysis filter bank for finite-length signals with N=24, N_0=20 and N_1=16 is shown in Fig. A.2, where N is the length of the input signal processed by Discrete Fourier Transform (DFT), N_0 is the length of the DFT of the low-pass filter signal after scaling, and N_1 is the length of the DFT of high-pass filter signal after scaling. The behavior includes two steps: filtering and scaling.

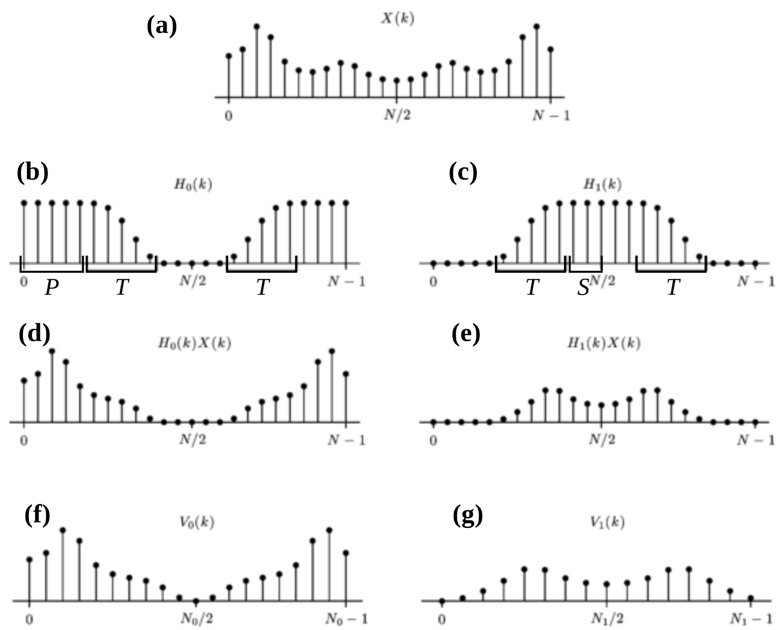

Fig. A.2 Behavior of the two-channel analysis filter bank for finite-length signals with N=24, N_0=20 and N_1=16. (a) DTF of input signal X(k), (b) low-pass filter $H_0(k)$, (c) high-pass filter $H_1(k)$, (d) DFT of input signal after low-pass filtering, (e) DFT of input signal after high-pass filtering, (f) DTF of low-pass filtered signal after scaling $V_0(k)$, and (g) DTF of high-pass filtered signal after scaling $V_1(k)$. (cited from [62])

The DFT of the N-point out signal is $y(n)$ is given by

$$Y(k) = \begin{cases} |H_0(k)|^2 X(k) & k \in P \\ \left(|H_0(k)|^2 + |H_1(k)|^2\right) X(k) & k \in T \\ |H_1(k)|^2 X(k) & k \in S \end{cases} \tag{A.7}$$

where the interval P refers to the pass-band of the low-pass filter $H_0(k)$ and the stop-band of the high-pass filter $H_1(k)$, the interval S refers to the pass-band of $H_1(k)$ and the stop-band of $H_0(k)$, and the interval T refers to transition-bands of the filter. The transition band uses the Daubechies filter with two-vanishing moments, given by

$$\theta(k) = \frac{1}{2}\left(1+\cos\left(\frac{k\pi}{T+1}\right)\right)\sqrt{2-\cos\left(\frac{k\pi}{T+1}\right)} \tag{A.8}$$

Pseudocode for the analysis filter bank used for TQWT is shown in Fig. A.3.

Function AFB(X, N_0, N_1)

Require: length(X), N_0 and N_1 even, N_0+N_1>length(x)

Output: V_0, V_1 (lengths N_0, N_1)

N=length(X)

P=(N - N_1)/2

T=(N_0+ N_1- N)/2-1

S=(N - N_0)/2

for $1 \le k \le T$ **do**

$\theta(k) = 0.5(1+\cos(k\pi/(T+1)))sqrt(2-\cos(k\pi/(T+1)))$

Low-pass subband:

$V_0(0)=X(0)$

for $1 \le k \le P$ **do**

$V_0(k)=X(k)$

$V_0(N_0-k)=X(N-k)$

for $1 \le k \le T$ **do**

$V_0(P+k)=X(P+k)\theta(k)$

$V_0(N_0-P-k)=X(N-P-k)\theta(k)$

$V_0(N_0/2)=0$

Hign-pass subband:

$V_1(0)=0$

for $1\leq k \leq T$ **do**

$V_1(k)=X(P+k)\theta(T+1-k)$

$V_1(N_1-k)=X(N-P-k)\theta(T+1-k)$

for $1\leq k \leq S$ **do**

$V_1(T+k)=X(P+T+k)$

$V_1(N_1-T-k)=X(N-P-T-k)$

$V_1(N_1/2)=X(N/2)$

Fig. A.3 Pseudocode for analysis filter bank (cited from [62])

With the analysis filter bank, the pseudocode for TQWT is shown in Fig. A.4.

Function TQWT(x, Q, r, P)

Require: length(x) even, $Q\geq 1$, $r>1$, P is positive integer

Output: wavelet coeffcients $\mathbf{w}^{(j)}$ and $\mathbf{w}^{(P+1)}$, $1\leq j \leq P$

$\beta= 2/(Q+1)$

$\alpha= 1-\beta/r$

$\mathbf{X}$=DFT(x)

N=length(x)

for $1\leq j \leq P$ **do**

$N_0 = 2\text{round}(\alpha^j N/2)$

$N_1 = 2\text{round}(\beta\alpha^{j-1} N/2)$

$(\mathbf{X}, \mathbf{W})=\text{AFB}(\mathbf{X}, N_0, N_1)$

$\mathbf{w}^{(j)}=\text{DFT}^{-1}(\mathbf{W})$

$\mathbf{w}^{(P+1)}= \text{DFT}^{-1}(\mathbf{X})$

Fig. A.4 Pseudocode for TQWT (cited from [62])

Similarly, to realize the inverse TQWT, a synthesis filter bank is needed. Pseudocode for the synthesis filter bank is shown in Fig. A.5.

Function SFB(V_0, V_1, N)

Require: V_0 and V_1 as in AFB, N even

Output: **Y** (length N)

N_0 =length(V_0)

N_1 = length(V_1)

P=(N - N_1)/2

T=(N_0+ N_1- N)/2-1

S=(N - N_0)/2

for $1 \leq k \leq T$ **do**

$\theta\ k = 0.5(1+\cos(k\pi/(T+1)))sqrt(2-\cos(k\pi/(T+1)))$

Low-pass subband:

$Y_0(0)= V_0(0)$

for $1 \leq k \leq P$ **do**

$Y_0(k)= V_0(k)$

$Y_0(N_0-k)= V_0(N-k)$

for $1 \leq k \leq T$ **do**

$Y_0(P+k)= V_0(P+k)\theta(k)$

$Y_0(N_0-P-k)= V_0(N-P-k)\theta(k)$

for $1 \leq k \leq S$ **do**

$Y_0(P+T+k)=0$

$Y_0(N-P-T-k)=0$

$Y_0(N/2)=0$

Hign-pass subband:

$Y_1(0)=0$

for $1 \leq k \leq P$ **do**

$Y_1(k)=0$

$Y_1(N_1-k)=0$

for $1 \leq k \leq T$ **do**

$Y_1(P+k)= V_1(k)\theta(T+1-k)$

$Y_1(N-P-k)=X(N_1-k)\ \theta(T+1-k)$

for $1\leq k \leq S$ **do**

$Y_1(P+T+k)=V_1(T+k)$

$Y_1(N-P-T-k)=V_1(N_1-T-k)$

$Y_1(N/2)=V_1(N_1/2)$

for $1\leq k \leq S$ **do**

$Y_1(k)= Y_0(k)+Y_1(k)$

Fig. A.5 Pseudocode for synthesis filter bank (cited from [62])

With the synthesis filter bank, pseudocode for inverse TQWT is shown in Fig. A.6.

Function iTQWT($\mathbf{w}^{(j)}$, $\mathbf{w}^{(P+1)}$, Q, r, P)

Require: Q, r, and P as in TQWT

Output: $\mathbf{y}$

$\beta= 2/(Q+1)$

$\alpha= 1-\beta/r$

$\mathbf{Y}$=DFT($\mathbf{w}^{(P+1)}$)

N=length(x)

for $j = P$ down to 1 **do**

$\mathbf{W}$ =DFT($\mathbf{w}^{(j)}$)

$M=2\text{round}(\alpha^{j-1}N/2)$

$\mathbf{Y}$=AFB($\mathbf{Y}$, $\mathbf{W}$, M)

$\mathbf{y}= \text{DFT}^{-1}(\mathbf{Y})$

Fig. A.6 Pseudocode for inverse TQWT (cited from [62])

Appendix B: Solution of the optimization problem by using the SALSA

The rationale of SALSA [86] is converting an unconstrained optimization problem with form given by

$$\min_{\mathbf{w}} f_1(\mathbf{w}) + f_2(\mathbf{w}) \tag{B.1}$$

into a constrained problem given as

$$\begin{gathered} \min_{\mathbf{u},\mathbf{w}} f_1(\mathbf{u}) + f_2(\mathbf{w}) \\ \text{such that } \mathbf{u} = \mathbf{w} \end{gathered} \tag{B.2}$$

The solution of the problem can be reached by an iterative algorithm composed of Eq. (B.3), Eq. (B.4) and Eq. (B.5):

$$\mathbf{u}^{k+1} = \arg\min_{\mathbf{u}} f_1(\mathbf{u}) + \mu \left\| \mathbf{u} - \mathbf{w}^k - \mathbf{d}^k \right\|_2^2 \tag{B.3}$$

$$\mathbf{w}^{k+1} = \arg\min_{\mathbf{w}} f_2(\mathbf{w}) + \mu \left\| \mathbf{u}^{k+1} - \mathbf{w} - \mathbf{d}^k \right\|_2^2 \tag{B.4}$$

$$\mathbf{d}^{k+1} = \mathbf{d}^k - \mathbf{u}^{k+1} + \mathbf{w}^{k+1} \tag{B.5}$$

To solve the problem given by Eq. (4.8), SALSA is applied with

$$f_1(\mathbf{u}) = \lambda_h \left\| \mathbf{u}_h \right\|_1 + \lambda_l \left\| \mathbf{u}_l \right\|_1 \tag{B.6}$$

$$f_2(\mathbf{w}) = \left\| \mathbf{y} - \mathbf{S}\mathbf{w} \right\|_2^2 \tag{B.7}$$

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_h \\ \mathbf{u}_l \end{bmatrix}, \mathbf{w} = \begin{bmatrix} \mathbf{w}_h \\ \mathbf{w}_l \end{bmatrix}, \mathbf{S} = \left[\mathbf{S}_h, \mathbf{S}_l \right] \tag{B.8}$$

Thus, the algorithm becomes

$$\mathbf{u}^{k+1} = \arg\min_{\mathbf{u}_h,\mathbf{u}_l} J(\mathbf{u}), J(\mathbf{u}) = \lambda_h \left\| \mathbf{u}_h \right\|_1 + \lambda_l \left\| \mathbf{u}_l \right\|_1 + \mu \left\| \mathbf{u}_h - \mathbf{w}_h^k - \mathbf{d}_h^k \right\|_2^2 + \mu \left\| \mathbf{u}_l - \mathbf{w}_l^k - \mathbf{d}_l^k \right\|_2^2 \tag{B.9}$$

$$\mathbf{w}^{k+1} = \arg\min_{\mathbf{w}} J(\mathbf{w}), J(\mathbf{w}) = \left\| \mathbf{y} - \mathbf{S}\mathbf{w} \right\|_2^2 + \mu \left\| \mathbf{u}^{k+1} - \mathbf{w} - \mathbf{d}^k \right\|_2^2 \tag{B.10}$$

$$\mathbf{d}^{k+1} = \mathbf{d}^k - \mathbf{u}^{k+1} + \mathbf{w}^{k+1} \tag{B.11}$$

Eq. (B.9) is then solved as a soft-threshold function by taking the partial differential of $J(\mathbf{u})$ about $\mathbf{u}$ [123]. The solution is given as

$$\mathbf{u}^{k+1} = soft\left(\mathbf{w}^{k} + \mathbf{d}^{k}, \frac{\lambda}{2\mu}\right) \tag{B.12}$$

Similarly, taking the partial differential of $J(\mathbf{w})$ about $\mathbf{w}$, Eq. (B.10) becomes

$$\mathbf{w}^{k+1} = \left(\mathbf{S}^{T}\mathbf{S} + \mu\mathbf{I}\right)^{-1}\left(\mathbf{S}^{T}\mathbf{y} + \mu\left(\mathbf{u}^{k+1} - \mathbf{d}^{k}\right)\right) \tag{B.13}$$

Since the TQWT is a tight frame, the bases satisfy $\mathbf{S}_h\mathbf{S}_h^T=\mathbf{I}$ and $\mathbf{S}_l\mathbf{S}_l^T=\mathbf{I}$ which implies $\mathbf{S}\mathbf{S}^T=2\mathbf{I}$. According to the Sherman-Morrison-Woodbury formula given by ([124],[125]),

$$\left(\mathbf{A} + \mathbf{m}\mathbf{n}^{T}\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{m}\mathbf{n}^{T}\mathbf{A}^{-1}}{1 + \mathbf{n}^{T}\mathbf{A}^{-1}\mathbf{m}} \tag{B.14}$$

Taking $\mathbf{A}$ as $\mu\mathbf{I}$, $\mathbf{m}$ as $\mathbf{S}^T$, $\mathbf{n}$ as $\mathbf{S}$ and substituting $\mathbf{S}\mathbf{S}^T=2\mathbf{I}$, we have

$$\left(\mathbf{S}^{T}\mathbf{S} + \mu\mathbf{I}\right)^{-1} = \frac{1}{\mu}\left(\mathbf{I} - \frac{1}{\mu + 2}\mathbf{S}^{T}\mathbf{S}\right) \tag{B.15}$$

Eq. (B.13) can be then expanded as

$$\mathbf{w}^{k+1} = \frac{1}{\mu + 2}\mathbf{S}^{T}\left(\mathbf{y} - \mathbf{S}\left(\mathbf{u}^{k+1} - \mathbf{d}^{k}\right)\right) + \left(\mathbf{u}^{k+1} - \mathbf{d}^{k}\right) \tag{B.16}$$

Hence, Eq. (B.11) is simplified to

$$\mathbf{d}^{k+1} = \frac{1}{\mu + 2}\mathbf{S}^{T}\left(\mathbf{y} - \mathbf{S}\left(\mathbf{u}^{k+1} - \mathbf{d}^{k}\right)\right) \tag{B.17}$$

By switching the order of Eq. (B.10) and Eq. (B.11), the iterative solution of the optimization problem via the SALSA is given by

$$\mathbf{u}_{i}^{k+1} = soft\left(\mathbf{w}_{i}^{k} + \mathbf{d}_{i}^{k}, \frac{\lambda_{i}}{2\mu}\right), i = h, l \tag{B.18}$$

$$\mathbf{b}_{i}^{k+1} = \mathbf{u}_{i}^{k+1} - \mathbf{d}_{i}^{k}, i = h, l \tag{B.19}$$

$$\mathbf{c}^{k+1} = \frac{1}{\mu+2}\left(\mathbf{y} - \mathbf{S}_h\mathbf{b}_h^{k+1} - \mathbf{S}_l\mathbf{b}_l^{k+1}\right) \tag{B.20}$$

$$\mathbf{d}_i^{k+1} = \mathbf{S}_i^T\mathbf{c}^{k+1}, i = h, l \tag{B.21}$$

$$\mathbf{w}_i^{k+1} = \mathbf{d}_i^{k+1} + \mathbf{b}_i^{k+1}, i = h, l \tag{B.22}$$

where $\mathbf{u}_i$, $\mathbf{b}_i$, $\mathbf{c}_i$ and $\mathbf{d}_i$ are all intermediary variables.