An Improved Empirical Wavelet Transform and Its Applications in Rolling Bearing Fault Diagnosis

Yonggang Xu *, Kun Zhang, Chaoyong Ma *, Xiaoqing Li and Jianyu Zhang

The Key Laboratory of Advanced Manufacturing Technology, Beijing University of Technology, Beijing 100124, China; zkun212@163.com (K.Z.); lixq_12@163.com (X.L.); zbyj_1999@bjut.edu.cn (J.Z.)
* Correspondence: xyg_1975@163.com (Y.X.); machaoyong@bjut.edu.cn (C.M.); Tel.: +86-010-6739-1750 (Y.X.)

Received: 16 October 2018; Accepted: 19 November 2018; Published: 22 November 2018

Abstract: As essential but easily damaged parts of rotating machinery, rolling bearings have been deeply researched and widely used in mechanical processes. The real-time detection of bearing state and simple, rapid, and accurate diagnosis of bearing fault are indispensable to the industrial system. The bearing’s inner ring and outer ring vibration acceleration can be measured by high-precision sensors, and the running state of the bearing can be effectively extracted. The empirical wavelet transform (EWT) can adaptively decompose the vibration acceleration signal into a series of empirical modes. However, this method not only runs slowly, but also causes inexplicable empirical modes due to the unreasonable boundaries of the frequency domain division. In this paper, a new method is proposed to improve the empirical wavelet transform by dividing the boundaries from the spectrum, named the fast empirical wavelet transform (FEWT). The proposed method chooses different points in the Fourier transform of the spectrum (key function) to reconstruct the trend component of the spectrum. The minimum points in the trend component divide the spectrum into a series of bands. A more reasonable set of boundaries can be found by choosing appropriate trend components to obtain effective empirical modes. The simulation results show that the proposed method is effective and that the acquired empirical mode is more reasonable than the EWT method. Combining kurtosis with fault feature extraction of inner and outer rings of bearings, the method is successfully applied to the fault diagnosis of rolling bearings.

Keywords: reasonable band; real-time detection; trend component; empirical wavelet transform; fault diagnosis

1. Introduction

Mechanical automation occupies a dominant position in modern industrial systems. Rotary machinery and bearings are indispensable to high-speed rail, intelligent robots, manipulators, and large-scale equipment production lines. In many cases, bearings may be subjected to high speed and a heavy load and may also face complicated conditions such as high and low-temperature differences. Even if the bearing structure is very delicate, it is difficult to tolerate a too-complex working environment for a long time. Sometimes minor damage can spread undetected, causing damage to the bearing and thus affecting the operation of the equipment. Healthy bearings are the basis for the normal operation of the equipment. Bearings that are prone to damage and breakdown are also one of the greatest hidden dangers that lead to equipment failure [1]. It can be seen that it is necessary to monitor and diagnose bearing faults.

With the popularization of mechanical automation and industrial intelligence, fewer people participate in the operation and management of equipment [2]. Real-time state-monitoring systems can compensate for the safety problems caused by manpower shortages. By monitoring the running state of mechanical equipment and collecting vibration acceleration signals of vulnerable parts, abnormal
equipment can be found in time. When rotating machinery fails, the position deformation of the fault will lead to repeated collisions and periodic impact [3]. Extracting the components containing the impact information in the signal can improve the speed and accuracy of the diagnosis. Many scholars have studied methods of diagnosing vibration acceleration signals, such as short-time Fourier transform (STFT), Wigner–Ville distribution, and wavelet transform (WT) [4,5]. The window function of STFT cannot be adaptively adjusted according to the characteristic frequency of the signal itself, Wigner–Ville distribution generates cross-terms when processing signals with multiple frequency components and wavelet analysis (including continuous wavelet transform, discrete wavelet transform, dual-tree complex wavelet transform, etc.) [4] requires manual selection of wavelet basis functions and lacks self-adaptability. Empirical mode decomposition (EMD), proposed by Huang [6], can adaptively decompose a signal into a finite number of intrinsic mode functions (IMFs). These contain local features of different time scales in the original signal. However, the problems of modal aliasing, endpoint effect, and over envelope seriously affect the further development of EMD [7,8]. In order to avoid modal aliasing and weaken the endpoint effect to a certain extent, Gilles [9] proposed extracting modal information from the spectrum via empirical wavelet transform (EWT). The method is adaptive and reliable in the mathematical derivation. First, the Fourier spectrum is divided adaptively. Then the wavelet filter banks are set up to extract multiple single components with tightly supported Fourier spectra. Finally, the Hilbert transform is used to get instantaneous frequency and amplitude for each component. With these advantages, the EWT method was used to extract fault information from rolling bearings [10,11] and wind turbines in [12]. Real-time monitoring of power quality has also been widely promoted [13]. By comparing EMD, Ensemble EMD, EWT, and other methods, Kedadouche [14] and others came to the conclusion that the EWT method can effectively reveal the frequency and associated harmonics of faulty bearings.

Some new problems were discovered in the process of using EWT to extract faults. For example, the setting of parameters is usually difficult and the calculation is time-consuming [15], the unreasonable division of the spectrum will lead to unsatisfactory noise reduction [16] or decomposition of too many invalid components [17], and so on. After consulting the literature, authors have divided the methods of optimizing EWT into three categories. The first category is processing the empirical modes (EMs) obtained. This includes removing noise and combining empirical modes by calculating relevant indicators. Empirical modes reduce their own noise through the selection of mutual information; the reconstructed signals are used to combine the ambiguity function and correlation coefficients for ambiguity correlation classification (ACC). Finally, new signals are input into the ACC for testing and training [16]. Pan [18] and Song [19] calculated the Pearson’s correlation coefficient of the envelope spectrum of each empirical mode. By comparing the coefficients of these modes and combining the empirical modes below the standard, a slightly smaller empirical mode can be obtained. This process combines noise modes and reduces the number of empirical modes. Mingtao [20,21] calculated root mean square, kurtosis, and skewness of each empirical mode and combined them into feature vectors; kernel density estimation (KDE) and mutual information (MI) are combined to estimate classification feature vectors to diagnose faults. Correlation coefficients among empirical modes are calculated to extract more valuable components; the instantaneous amplitude was extracted by HT and STFT and the damage was characterized [22]. In view of the inconvenience of EWT spectrum segmentation caused by noise, Fourier–Bessel series expansion (FBSE) is mentioned to enhance spectrum and reduce noise; the scale-space method is used to split the FBSE spectrum [23]. This method suppresses the noise to some extent, but does not solve the problems of mode aliasing and invalid components caused by the scale–space method in EWT. Similarly, Luo [24] used the autoregressive power spectrum instead of the Fourier spectrum to divide boundaries; then, the boundaries were computed by the Burg algorithm to correct EWT. This approach also enhances the spectrum and reduces noise, but does not explain how to divide the boundaries in signals containing modulation components. This category adds more work on the premise that EWT takes more time. It takes a lot of time not only to set the initial parameters of EWT, but also to
set the merge criteria. This will only make the algorithm more complicated. Although the noise is reduced to some extent, or the number of empirical modes is reduced, it does not change the defect of the EWT method itself. Therefore, it is necessary to improve the algorithm. The second category is optimizing the adaptability of EWT. Due to the disturbance of too many initial parameters, Zheng \[15\] proposed an adaptive parameterless empirical wavelet transform method. This method uses Otsu and K-means to segment the spectrum and establishes a set of adaptive bandpass filters to decompose the signal into mono-components with a compactly supported spectrum. Another adaptive method uses sparsity functions, first prewhitening the signal by using an autoregressive model, then automatically determining Fourier segments to conduct EWT and square envelope analysis \[25\]. Otsu and K-means methods have poor effects on spectrum segmentation; sparsity guide EWT needs to change parameters according to fault characteristics while optimizing parameters. The third category is optimizing spectrum segmentation. Hu \[17,26\] proposed a new method of spectrum segmentation, which uses an order statistical filter (OSF) to estimate the upper envelope function of the spectrum; by selecting the useful flat top in the upper envelope, the boundaries are determined to construct the filter and reconstruct the empirical modes. OSF is very effective in estimating envelope function, but the author did not point out how to get the best envelope function. Therefore, the method is not self-adaptive. Juan \[27\] used multiple signal classification to calculate the spectrum, which zeroes the part considered as noise and enlarges the part deemed useful. On the one hand, using multiple signal classification to calculate the spectrum of signals can reduce noise; on the other hand, it can amplify useful information while weakening the irrelevant information, so as to facilitate the division. In Reference \[28\], the natural frequencies are used as useful frequencies to calculate the support boundaries and decompose the signal into multiple components; kurtosis is used to automatically filter defect-related components and envelopes to extract characteristic frequency. However, the above three methods have certain limitations: the noise interference is very obvious when calculating the signal spectrum, and how to distinguish the strong noise from the signal needs to be further explored. Calculating the cutoff frequency and initializing the boundary with the type of fault needs to be performed; the final boundaries are determined by an iterative backward forward search algorithm, and MFMindex-aided adaptive Fourier spectrum segmentation is realized \[29\]. The multilevel EWT method proposed by Ding \[30\] simplifies the rules for splitting the spectrum. A multilevel boundary array is constructed based on the rotational frequency, outer race fault characteristic frequency, inner race fault characteristic frequency, roller fault characteristic frequency, and cage fault characteristic frequency and used for filtering and reconstruction. This method provides a complete framework for splitting the spectrum without depending on the characteristics of the spectrum. However, because of the difference between roller fault and inner ring fault characteristic frequencies, the decomposition level will be different. Self-adaptability is improved, while computing time is wasted. More invalid modes are extracted from each level. Therefore, optimizing the core of the EWT algorithm—segmentation of the spectrum—has become the focus of this research.

In this paper, a fast empirical wavelet transform is proposed. Based on the waveform characteristics of the spectrum, the problem of spectrum segmentation in EWT is improved. The trend of the spectrum is reconstructed from the Fourier function of the spectrum, and the minimum points of the trend are calculated as the boundaries. Different frequency components in the key function can reconstruct different trend components. By comparing the success rate and error rate of separation, it is proved that the boundaries of the fast empirical wavelet transform (FEWT) method are more reasonable. The experimental results show that the method is effective in the analysis of rolling bearing fault diagnosis. The method of dividing boundaries based on the Fourier transform function of the spectrum is not only fast but also self-adaptive, and can be effectively applied to online fault diagnosis systems. The structure of this paper is as follows: Section 2 briefly introduces EWT and its shortcomings; Section 3 elaborates the proposed FEWT method in detail and introduces the trend estimation method based on key function and threshold denoising; Section 4 uses three kinds of simulation signals to prove that the trend estimation method can obtain more reasonable boundaries
and extract useful information effectively; and Section 5 combines kurtosis and fast kurtogram to verify the validity of the method through the fault of outer and inner rings of rolling bearings.

2. Exposition of Traditional EWT Method

2.1. EWT Method

The EWT method can be described by the following three steps: (1) adaptively segment the spectrum; (2) construct a suitable empirical wavelet filter bank according to the boundaries and filter the signal; and (3) reconstruct empirical modes and use Hilbert transform to demodulate and analyze.

Step 1: Divide the boundaries in the frequency domain.

The frequency domain of the signal is normalized to $[0, \pi]$ and divided into $N$ frequency bands with different bandwidths. Let $\omega$ be the boundary between the bands; the left boundary is $\omega_0 = 0$, the right boundary is $\omega_N = \pi$, and the remaining boundaries can be expressed as $\omega_1, \omega_2, \ldots, \omega_N$, then each frequency band can be recorded as $\Lambda_n = [\omega_{n-1}, \omega_n], n = 1, 2, \ldots, N$. It is obvious that $\bigcup_{n=1}^N \Lambda_n = [0, \pi]$. With each boundary as the center, a transition phase with a width of $T_n = 2\tau_n$ is defined for building a window base.

In this paper, the spectral trend estimation method based on the key function is used to obtain the boundary of each frequency band; instead of the method used by Gilles, the midpoint of the adjacent maxima is used as the boundary condition, as shown in Section 3.1.

Step 2: Construct an empirical wavelet filter bank.

The Meyer wavelet was selected by Gilles as a basis function to construct an empirical wavelet. A group of mutually orthogonal trigonometric functions is designed in the transition phase where the boundary of the frequency band is located, and a constant is designed in the frequency band. The empirical scaling function $\hat{\varphi}_n(\omega)$ and the empirical wavelets $\Psi_n(\omega)$ can be defined as:

$$
\hat{\varphi}_n(\omega) = \begin{cases} 
1; & |\omega| \leq (1 - \gamma)\omega_n \\
\cos\left[\frac{\pi}{2} \beta \left(\frac{1}{\omega}\right) \left(|\omega| - (1 - \gamma)\omega_n\right)\right] ; & (1 - \gamma)\omega_n \leq |\omega| \leq (1 + \gamma)\omega_n \\
0; & \text{others}
\end{cases}
$$

$$
\Psi_n(\omega) = \begin{cases} 
1; & (1 + \gamma)\omega_n \leq |\omega| \leq (1 - \gamma)\omega_{n+1} \\
\cos\left[\frac{\pi}{2} \beta \left(\frac{1}{\omega}\right) \left(|\omega| - (1 - \gamma)\omega_{n+1}\right)\right] ; & (1 - \gamma)\omega_{n+1} \leq |\omega| \leq (1 + \gamma)\omega_n \\
\sin\left[\frac{\pi}{2} \beta \left(\frac{1}{\omega}\right) \left(|\omega| - (1 - \gamma)\omega_{n+1}\right)\right] ; & (1 - \gamma)\omega_n \leq |\omega| \leq (1 + \gamma)\omega_{n+1} \\
0; & \text{others}
\end{cases}
$$

where the transition function $\beta(x)$, the coefficient $\gamma$, and the transition phase $\tau_n$ are

$$
\beta(x) = x^4 \left(35 - 84x + 70x^2 - 20x^3\right)
$$

$$
\gamma < \min\left(\frac{\omega_{n+1} - \omega_n}{\omega_{n+1} + \omega_n}\right)
$$

$$
\tau_n = \gamma \omega_n , \quad 0 < \gamma < 1
$$

Step 3: Conduct an empirical wavelet transform.

The empirical wavelet transform $W_f$ is defined by a method similar to the traditional wavelet transform, the Fourier transform is $F(\cdot)$, and the inverse Fourier transform is $F^{-1}(\cdot)$.

$$
W_f(n, t) = \langle f(t), \Psi_n(t) \rangle = \int f(\tau) \overline{\Psi_n(\tau - t)} d\tau = F^{-1}(f(\omega) \hat{\Psi}_n(\omega))
$$
Generate the approximation coefficients $W_\varepsilon f(n, t)$ by the inner product of the signal with the empirical scaling function:

$$W_\varepsilon f(0, t) = \langle f(t), \varnothing_1(t) \rangle = \int f(\tau) \varnothing_1(t - \tau) d\tau = F^{-1} (f(\omega) \hat{\varnothing}_1(\omega))$$

(7)

where $\mathbf{x}(\omega)$, $\hat{\varnothing}_1(\omega)$, $\hat{\Psi}_n(\omega)$ can be expressed as Fourier transforms of $x$, $\varnothing_1(\omega)$, and $\Psi_n(\omega)$. The signal can be rebuilt as follows:

$$f(t) = W_\varepsilon f(0, t) * \varnothing_1(t) + \sum_{n=1}^{N} W_\varepsilon f(n, t) * \Psi_n(t)$$

(8)

where $\hat{W}_\varepsilon f(0, \omega)$ and $\hat{W}_\varepsilon f(n, \omega)$ can be expressed as Fourier transforms of $W_\varepsilon f(0, t)$ and $W_\varepsilon f(n, t)$. The empirical mode can be given by

$$\begin{align*}
  f_0(t) &= W_\varepsilon f(0, t) * \varnothing_1(t) \\
  f_k(t) &= W_\varepsilon f(n, t) * \Psi_k(t)
\end{align*}$$

(9)

After obtaining the empirical mode of the signal, the Hilbert transform can be used to calculate instantaneous frequency and instantaneous amplitude or to diagnose faults.

### 2.2. Drawbacks of the EWT Method

The key step in the process of an empirical wavelet transform is dividing the appropriate boundaries in the spectrum. When a large number of parameters are set up to achieve adaptive partitioning, it is difficult in most cases to obtain the appropriate boundary: (1) it is difficult to separate them when the spectra of different components are closer, which will lead to modal aliasing; and (2) when the signal-to-noise ratio (SNR) is low, the noise components cannot be eliminated and more invalid components will be obtained.

In Figure 1, the areas labeled A show that the two cosine signals with a relatively close frequency are not separated, and the area labeled B shows that the noise is divided into many ineffective components.

![Figure 1](image.png)

**Figure 1.** The spectrum segmentation of the empirical wavelet transform (EWT) method: (a) adaptive; (b) scale-space empirical law type.

The adaptive method needs to set parameters: InitBounds. The initial boundary is difficult to determine, and the number of edges directly affects the results of segmentation. There are a number of types of split spectrum in scale-space, but they lead to too many invalid components in the frequency domain, as shown in B. Therefore, it is necessary to explore a fast and reasonable method to divide the boundaries from the spectrum to improve the empirical wavelet transform.
Figure 2. A flowchart of the fast empirical wavelet transform (FEWT) method.
3. Proposed Method of Fast Empirical Wavelet Transform

In this paper, the above two defects of EWT are discussed and expounded: spectrum segmentation and boundary optimization. A new improved empirical wavelet transform method is proposed. This method obtains the trend component of the spectrum by obtaining the Fourier transform function (key function) of the spectrum. Minimum points of the trend component are used as boundaries to reconstruct the signal. Selecting different points in key functions will reconstruct different trend components. Suitable trend components can obtain reasonable boundaries for EWT. Figure 2 shows the process of the proposed method. The specific steps are as follows:

Step 1: Collect periodic impact signals \(y(t)\). The fast Fourier transform (FFT) algorithm is used to acquire spectrum \(Y(f)\). The Fourier transform function of \(Y(f)\) is obtained through the FFT algorithm again, namely, the key function \(K(f)\).

Step 2: A certain number of points in the key function are extracted to inverse the Fourier transform and reconstruct the trend component \(T_c(f)\) of the spectrum. When more points are used to reconstruct the complex trend component, the number of extreme points of the trend component will be larger, so that the spectrum will be divided into many parts. The reconstructed function is close to the trend of spectrum changes, and the spectrum will be partitioned into fewer parts. Section 3.1 describes the details of this method.

Step 3: Calculate the boundaries of the spectrum and optimize.

Taking the trend component as the basis function, the minimum point sequence is used as the boundary array. When the trend component is complex, the boundary array is optimized by the wavelet threshold. The method of optimizing spectrum boundaries based on threshold denoising reduces the invalid components caused by noise to a certain extent. This method will be introduced in Section 3.2.

Step 4: Perform signal processing using empirical wavelet transform.

The frequency domain of the signal is normalized to \([0, \pi]\); choose Meyer wavelet as the basis function; define the empirical scaling function \(\varnothing_n(\omega)\) and the empirical wavelets \(\Psi_n(\omega)\); build adaptive filters; decompose the signal into different components.

3.1. Trend Estimation Method Based on the Key Function

In the proposed method, the step of trend estimation based on the key function can be used to reconstruct the trend of the spectrum. This method inherits the theoretical basis of the Fourier transform. It is better than an interpolation algorithm in terms of computation time; the algorithm is more effective than the traditional empirical wavelet transform. Figure 3 shows the process of obtaining trend components.

Step 1: Obtain the spectrum of the measured signal using the FFT algorithm, as shown in Figure 3a.

\[ Y(f) = FFT(y(t)) \] (10)

Step 2: Calculate the Fourier transform function of \(Y(f)\). Figure 3b,d show the key function.

\[ K(f) = FFT(Y(f)) \] (11)

Step 3: The components of key functions are closely related to the spectrum. Different functions can be obtained through low-pass filtering of key functions. The blue line in Figure 3c can be obtained by inverse FFT (iFFT) low-pass filtering with \(B_1 = 30\) as the boundary. This blue line approximates the trend of the spectrum, called the trend component:

\[ T_c(f) = iFFT(K_B(f)) \] (12)

where \(K_B(f)\) is the left part of boundary \(B\) in \(K(f)\) and \(T_c(f)\) is a trend component.
The purple line in Figure 3e can be obtained by inverse FFT low-pass filtering with $B_1 = 300$ as the boundary. Compared with the above results, it is obvious that different trend components can be reconstructed by inverse FFT with different frequency components. Therefore, it is necessary to study how the number of reconstructed points affects trend components.

**Figure 3.** A flowchart for calculating trend components: (a) spectrum of the signal; (b) key function of the spectrum with $B_1 = 30$; (c) trend component of the key function with $B_1 = 30$; (d) key function of the spectrum with $B_2 = 300$; (e) trend component of the key function with $B_2 = 300$.

It can be seen that as the number of reconstructed points increases, the trend components become more and more complex; different trend components will have different division results. When the number of reconstructed points is smaller than the first peak value of the key function, such as $B_1$, the trend component is gentle and fluctuates very little, approximates a straight line, and cannot clearly be seen from the spectrum fluctuation trend. One minimum divides the spectrum into two areas. The first area contains a large amount of information and does not separate the corresponding frequency components in the spectrum. Thirty points are used to reconstruct the trend component (blue line), where the minimum values are regarded as the boundaries (black line). The key clusters in the spectrum are successfully extracted. Compared with the red line, the trend component corresponding to the blue line is divided into more bands, and the spectrum information contained in each band is more representative and readable. If the number of reconstructed points is increased to $B_5$, $B_6$, $B_7$, the trend component will become more complex and the number of minimum points will increase rapidly. The boundary divides the spectrum into more ineffective components. Comparing the seven trend components, the trend component can respond to the trend of spectral fluctuation when the number of reconstructed points is $B_3 = 30$. In order to accurately determine the optimal frequency component in the key function, the intervals of five frequency components
are $B_1 = 20$, $B_2 = 25$, $B_3 = 30$, $B_4 = 35$, $B_5 = 40$. To calculate trend components, as in Figure 4b, on the premise of 20 points spanning the front and back, the trend components are almost the same; the boundary sequence of trend minima is also similar. Therefore, it is meaningless to find a more accurate frequency component; it allows the number of reconstructed points to fluctuate in a small range.

Figure 4. (a) The trend components and boundaries of different reconstruction points after inverse fast Fourier transform (iFFT): $B_1 = 4$, $B_2 = 10$, $B_3 = 30$, $B_4 = 60$, $B_5 = 100$, $B_6 = 200$, $B_7 = 300$; (b) trend components and boundaries of different reconstruction points after iFFT: $B_1 = 20$, $B_2 = 25$, $B_3 = 30$, $B_4 = 35$, $B_5 = 40$. 
For signals with modulated components and bandwidth, it is necessary to estimate the trend to extract signals. The number of reconstructed points should be \( B \in [10, 60] \). In Figure 1, for two signals whose spectra are close to each other, it is necessary to divide them accurately in order to extract the signals. Rolling bearings are one of the most easily damaged parts in rotating machinery. When a rolling bearing fails, the rolling body will impact periodically when it touches the fault position [31,32]. Shock information exists with modulating components but does not exist alone. The impact and modulation components show a cluster of modulation information containing a side band in the spectrum. How to extract the modulation information of each cluster becomes the most critical step in the process of splitting the spectrum. The vibration acceleration sensor is sensitive to impact information. Trend components represent the trend of spectrum change, and can better fit the similar modulation information, which may contain shock information in each cluster of the spectrum. Therefore, the trend estimation method based on the key function has advantages in dealing with rolling bearing fault diagnosis. The method based on FFT and iFFT is fast in calculating trend components. It has obvious advantages over the traditional EWT in dealing with analog signals (see Section 4) and practical signals (see Section 5). Choosing too few points from the key function for the inverse transformation will lead to rough partition that cannot distinguish the information with close frequency, but it can well contain the modulation information with side-band frequency. The proposed method can preserve and exclude the useless components to a certain extent. For simulation signals, when many points are selected for inverse transformation, some details can be extracted, such as the components whose spectra are close to each other, but this will cause the spectrum to be divided into too many areas. This paper proposes a threshold noise-reduction method to optimize the above problems.

### 3.2. Threshold Denoising Method

Threshold denoising is mostly used for wavelet threshold denoising. After obtaining the wavelet coefficients through wavelet decomposition, the coefficients are denoised and the signals are reconstructed. In the process of threshold denoising, different threshold functions affect not only the denoising effect, but also the reconstruction results. It is critical to set the threshold when estimating data. Because this method does not need to show the trend completely, the spectrum fluctuation caused by noise needs to be reduced in the process of threshold denoising. Therefore, there is a certain fault tolerance when thresholds are obtained.

Commonly used threshold functions include the hard threshold function:

\[
\hat{s} = \begin{cases} 
    s \mid s \mid \geq \lambda \\
    0 \mid s \mid < \lambda 
\end{cases}
\]  

and the soft threshold function:

\[
\hat{s} = \begin{cases} 
    \text{sgn}(s) \cdot (|s| - \lambda) \mid s \mid \geq \lambda \\
    0 \mid s \mid < \lambda 
\end{cases}
\]  

where \( s \) is the signal to be processed and \( \hat{s} \) is the signal after denoising.

Donoho and Johnstone [33] give the fixed threshold of \( \lambda \):

\[
\lambda = \sigma \sqrt{2 \ln(N)}
\]  

where \( N \) is the length of the signal to be processed and \( \sigma^2 \) is the estimated noise variance:

\[
\sigma = \text{Mean}(s) / 0.6745
\]
The hard threshold can preserve the local information of the signal very well in the processing, but the discontinuity of \( s \) at \( \lambda \) and \(-\lambda\) leads to oscillation of the reconstructed signal at the discontinuity point. The overall continuity of \( \hat{s} \) is better during soft threshold processing. Although the derivatives are discontinuous and there is always a constant deviation between \( s \) and \( \hat{s} \), which may affect the accuracy of reconstruction, this method only needs to obtain the spectrum trend and can ignore these defects of the soft threshold. Figure 5 shows the differences in noise reduction between the hard and soft threshold, where the black line is a signal sequence. The signal sequence is processed by using \( \lambda = 0.5 \text{ m/s}^2 \) as the threshold. The red line is the result of a hard threshold; the blue line is the result of a soft threshold. It can clearly be seen from the graph that soft thresholds have a good continuity, while hard thresholds have local steps. In this paper, soft thresholds are used to process trend components.

![Figure 5. The traditional threshold functions.](image)

4. Simulated Signal Verification

EWT needs to set many parameters when dividing boundaries, and different parameters will get different results. This affects its adaptability to some extent. The time required to repeatedly modify a large number of parameters exceeds the time saved by its self-adaptation. Sometimes it is difficult to distinguish two components whose frequencies are close. In a low signal-to-noise ratio signal, EWT divides more invalid components to bring unnecessary trouble to the analysis. In view of the above problems, three sets of simulation signals are given in this section to prove that FEWT boundary demarcation is more reasonable, faster, and more efficient than EWT.

4.1. Case Study 1

The first simulation signal consists of three cosine signals with different frequencies and amplitudes.

\[
\begin{align*}
    s_1 &= 2.0 \cos(2\pi f_1 t) \\
    s_2 &= 1.5 \cos(2\pi f_2 t) \\
    s_3 &= 2.5 \cos(2\pi f_3 t) \\
    s &= s_1 + s_2 + s_3 + \eta
\end{align*}
\]

Among them, \( f_1 = 100 \text{ Hz}; f_2 = 150 \text{ Hz}; f_3 = 300 \text{ Hz}; \) and \( \eta = \text{SNR}(3) \).

The composition, noise, and spectrum of the simulation signal are shown in Figure 6. The signal contains three frequency components: 100 Hz, 150 Hz, and 300 Hz. The frequencies of the first two cosine signals are close to each other. How to successfully extract these three components and minimize the noise interference in the extracted components becomes an important problem to be solved.
First, the proposed EWT and FEWT methods are used to process the signals. Figure 7 shows the effect comparison of the two methods in terms of boundary segmentation. The long red dashed lines are the boundaries of FEWT and the short blue dashed lines are the boundaries of EWT. It can be seen from the figure that the blue boundaries divide the spectrum into too many areas. Although the components represented by $f_3$ were successfully extracted, $f_1$ and $f_2$ were not distinguished. This is one of the biggest defects of EWT. Another major defect, dividing too many regions, will lead to the same amount of invalid components in the next reconstruction, which may increase the workload of the subsequent analysis and interfere with diagnosis. In the process of signal processing by the FEWT method, the first 180 points are selected from key functions to reconstruct trend components. After calculation, the soft threshold denoising process is $\lambda = 0.2309$. The minimum point in the trend component after noise reduction can be well separated into three frequencies and contains only a few unavoidable noises. In Figure 7, the regions divided by long red dashed lines are more reasonable than the short blue dotted lines.

Since EWT decomposes too many components, only the first three components of the EWT containing all three simulation signals are extracted for comparison. As shown in Figure 8a, the EWT method does not separate the components of 100 Hz, 150 Hz, and 300 Hz. There is modal aliasing in EM1, and EM2 is the component with a frequency of 300 Hz. However, because it contains more noise, it causes a very unstable modulation phenomenon. EM3 and the remaining components are invalid components that only contain noise. In the process of processing the signal by FEWT, under the effect of threshold noise reduction, the spectral boundary shrinks closer to the frequency of the component. The components representing the 2, 4, and 6 frequency domains are shown in Figure 8b. It can be seen from the figure that the three components are perfectly separated. The amplitudes of the three components have only a small fluctuation, which proves that the noise contained in the signal is small. It is proved that the proposed method is suitable for signals with relatively close frequencies. At this point, the FEWT method is superior to the EWT method.
For the simulation signal, this paper verifies the rationality of the FEWT boundary from two aspects: the success rate of separation and the error rate. The formula for the success rate of separation is as follows:

\[
\text{rate\_Success} = \frac{\text{effective\,(EM)}}{\text{all\,(EM)}}
\]

where rate\_Success is the success rate of separation, effective\,(EM) is the number of empirical modes successfully separated, and all\,(EM) is the total number of empirical modes. After calculating the difference between the separated empirical mode and the corresponding component of the original signal, the mean square error is used to calculate the error rate. The formula is as follows:

\[
\text{rate\_Error} = \sqrt{\frac{\sum_{i=1}^{N} (\text{EM}_i - s_k)^2}{N}}
\]

where rate\_Error is the error rate, EM\(_j\) is the jth empirical model waiting for comparison, and s\(_k\) is the corresponding empirical mode.

Figure 9 shows the separation success rate and error rate of EWT and FEWT. The success rate of EWT decomposition is 7.14% and FEWT is 42.86%. The FEWT decomposition efficiency and success rate are higher. Since the EWT method only successfully separates the third component of the original signal, only the error rate of this component can be compared. From Figure 9b, it is known that the components decomposed by FEWT are more consistent with the original components. Therefore, the boundary of the proposed method is more reasonable.

**Figure 9.** (a) The success rate of separation; (b) error rate.

### 4.2. Case Study 2

The second signal simulates the working environment when the equipment breaks down. First, the periodic shock of equipment failure is simulated: a set of damped impulse signals are established and the inherent frequency of the impulse signals \(f_{n1} = 1000\) Hz, the damping coefficient \(g = 0.007\), and the repeated cycle \(T = 0.01\) s. So the characteristic frequency \(f_q = 100\) Hz. Second, the nonimpact and nonfault modulation signals \(s_2\) in rotating machinery operation are simulated. The inherent
frequency of this amplitude-modulated and frequency-modulated (AM-FM) signal is \( f_{n1} = 3000 \) Hz, the width of the side band is 100 Hz, and the noise is \( \eta = \text{SNR}(-0.03 \text{ db}) \).

\[
\left\{ \begin{array}{l}
s_1 = 7e^{-g \times 2\pi f_{n1} t} \times \sin\left(2\pi f_{n1} t \times \sqrt{1 - g^2}\right) \\
s_2 = 2 \sin(100\pi t) \sin(2\pi f_{n2} t + \sin(200\pi t)) + 0.8 \cos(4\pi t) \cos(100\pi t)
\end{array} \right. 
\tag{20}
\]

Figure 10a,b shows the components of the signal. The signal-to-noise ratio signal is shown in Figure 10c and its spectrum in Figure 10d. The central frequencies of the two components are far apart in the spectrum. Theoretically, EWT can distinguish these two components, but the difficulty in dividing boundaries of this kind of signal is reducing the noise interference and invalid components.

\[s = s_1 + s_2 + \eta\]

![Figure 10](image.png)

Figure 10. (a) The pure signal; (b) the amplitude-modulated and frequency-modulated (AM-FM) signal; (c) the signal-to-noise ratio (SNR) signal; (d) the spectrum.

In this paper, two methods are used to process the signal and the following results can be obtained. Figure 11a is the result of EWT. The impulse signal with an inherent frequency of 1000 Hz and the modulation signal with an inherent frequency of 3000 Hz are successfully separated. However, the following problems are also left: part A in Figure 11a has a lot of noise left; part B divides up too many invalid components.

![Figure 11](image.png)

Figure 11. The boundaries divided by (a) the EWT method and (b) the FEWT method (green line: trend component).
Figure 11b shows the result of the FEWT method. The first 30 points of the key function are selected to reconstruct the trend component. The threshold $\lambda = 0.0833$. After calculation, FEWT divides the frequency domain into seven parts, five fewer than EWT. Part A, which contains more noise in the result of EWT, is divided by the FEWT method. The FEWT method obtains fewer components than EWT. The reduction is shown in part B. Since part B is mostly noise, FEWT is more reasonable than EWT.

Two parts representing pulse information and modulation information are extracted: for the result of EWT, the third and eighth components are extracted; for the result of FEWT, the fourth and sixth components are extracted, as shown in Figure 12. Figure 12a,b show pulse and modulation information obtained by the EWT method, and Figure 12c,d by FEWT. Because the boundaries of the two methods are similar, the reconstructed components are almost the same, but the details are different. The modulation signal decomposed by EWT contains a lot of noise, which is segmented by EWT. Enlarging the data between 0.04 s and 0.06 s at part A in Figure 12b,d, such as Figure 12e,f, abnormal information can be found in the details of the modulated signal decomposed by EWT, as shown in part B. A similar signal is also found at the corresponding position C of the modulated signal decomposed by FEWT, but the amplitude is relatively small. It can be seen that FEWT can not only divide fewer and more reasonable modal components than EWT, but also reduce certain noise effects when processing pulse signals with modulated signals. The simulation results show that the FEWT method is superior to the EWT method.

Figure 13 shows the separation success rate and error rate of EWT and FEWT. The success rate of EWT decomposition is 16.67% and of FEWT is 28.58%. The FEWT decomposition efficiency and success rate are higher. The impact component (EM1) and modulation component (EM2) are compared and the results are shown in Figure 13b. It is known that the components decomposed by FEWT are more consistent with the original components. Therefore, the boundary of the proposed method is more reasonable.

![Figure 12](image12.jpg)

**Figure 12.** The modal components and details obtained by two methods. Pulse component decomposed by (a) the EWT method and (c) the FEWT method. Modulating component decomposed by (b) the EWT method and (d) the FEWT method. (e) Details of (b); (f) details of (d).

![Figure 13](image13.jpg)

**Figure 13.** (a) The success rate of separation; (b) error rate.
4.3. Case Study 3

The third simulation signal adds a periodic high-frequency impact to a low-frequency cosine signal locally. Periodic high-frequency $s_1$ impact the signals and simulates faults; low-frequency signals $s_2$ simulate the normal operation of equipment:

$$
\begin{align*}
  s_1 &= 2 \sin(2\pi 200t) \times (u(t - 0.09) - u(t - 0.15) + u(t - 0.34) - u(t - 0.4) + \\
       &\quad u(t - 0.59) - u(t - 0.65) + u(t - 0.84) - u(t - 0.9)) \\
  s_2 &= 10 \cos(2\pi 4t) \\
  s &= s_1 + s_2 + \eta
\end{align*}
$$

(21)

where $u(t)$ is the step signal, the sampling frequency is 2 kHz, collection time $t = 1$ s, and $\eta = \text{SNR}(3)$. The original signal and its constituents are shown in Figure 14. Figure 14a is the original signal that adds noise after superposition; b and c are constituents of the original signal.

![Figure 14](image)

**Figure 14.** (a) The original signal; (b) the cosine component; (c) the impact component.

EWT and FEWT were used to process the signals. In the FEWT calculation, the first 60 points of the key function are selected to reconstruct trend components. The threshold $\lambda = 0.0614$. Both methods successfully segment the low-frequency components. The difference lies in the location of the first boundary. The first boundary of the EWT method is closer at 100 Hz, and the FEWT is closer at 10 Hz. A comparison between the two results shows that FEWT is closer to 4 Hz, which results in less noise in the first component of the reconstruction. Figure 15a shows that the spectrum is divided into 14 regions by the EWT method, while FEWT divides the spectrum into only 4 regions. The ineffective regions caused by noise at part B are divided into independent components. This is not accurate. The excessive invalid component not only increases the computational cost and operation time, but also brings a difficulty to the diagnosis.

In Figure 16, the red line shows the constituents of the original signal; in Figure 16a, the black line is the first 4 of 14 components decomposed by the EWT; Figure 16b is all results of FEWT decomposition. It can be seen from the figure that both the low-frequency and high-frequency components of EWT decomposition contain certain noise. The low-frequency component of FEWT is very smooth. The high-frequency component of FEWT decomposition has a weak low-frequency modulation component but little noise.
Figure 15. The spectrum decomposition modes of (a) the EWT method and (b) the FEWT method.

Figure 16. (a) The first 4 of 14 components of EWT decomposition modes; (b) all components of FEWT decomposition modes.

Figure 17 shows the separation success rate and error rate of EWT and FEWT. The success rate of EWT decomposition is 14.29% and FEWT is 50%. FEWT decomposition efficiency and success rate are higher. The cosine component (EM1) and impact component (EM2) are compared, and the results are shown in Figure 17b. It is known that the components decomposed by FEWT are more consistent with the original components. Therefore, the boundary of the proposed method is more reasonable.

Figure 17. (a) The success rate of separation; (b) the error rate.
5. Applications

Rolling bearings are liable to be damaged during operation. It is necessary to collect bearing vibration data in real time with appropriate sensors to monitor the bearing state so as to avoid equipment failure. Two groups of experimental data are used to verify the method mentioned in this paper. The first group is the fault data of bearing outer ring, which verifies that this method is more accurate, rapid, and convenient than EWT when dealing with the fault. The second group is the failure data of bearing inner ring, which verifies the application of this method in fast kurtogram (FK) and the processing result is better.

5.1. Analysis of Bearing Outer Ring Fault Data

This experiment uses the fault bearing test data collected by the laboratory at Xi’an Jiaotong University. The Spectra Quest, Inc test bench is shown in Figure 18. The actual motor frequency is 33.6 Hz, the motor speed is 2000.9 r/min, and the sampling frequency is $F_s = 15,000$ Hz. The sensor is connected to the outer ring of the motor end bearing. After calculation, the fault characteristic frequency of the inner ring of the bearing can be obtained as $f_q = 102.4$ Hz.

![Figure 18. The Spectra Quest, Inc. test bench.](image)

Figure 19a shows the original signal and its spectrum, and Figure 19b is the envelope. There is no obvious periodic shock component in the original signal. In the envelope, there is neither a distinct peak nor a frequency multiplier with the same interval. It is difficult to find useful information to diagnose a fault.

![Figure 19. (a) The original signal and its spectrum; (b) the envelope of the signal.](image)
EWT is used to process signals, and the segmentation results are shown in Figure 20. The blue dotted line is the effect of spectrum segmentation based on the EWT method. As can be seen from the graph, this method divides many boundaries and gets a lot of empirical modes (EMs), which brings a lot of work to the fault analysis. The kurtosis of each EM is extracted and a broken line graph is drawn.

![Figure 20](image)

**Figure 20.** The effect of spectrum segmentation based on the EWT method (blue lines); kurtosis of each EM (red line).

The four largest components of kurtosis, A, B, C, and D, are extracted and their waveforms and envelopes are shown in Figure 21. In the envelope of A, B, and C with large kurtosis, the characteristic frequency with a prominent peak value and its frequency doubling cannot be found. Only component D has a characteristic frequency but no frequency doubling information.

![Figure 21](image)

**Figure 21.** The EM7, EM11, EM12, and EM23 and their envelopes.

FEWT is used to process data, with 45 points of the key function selected to reconstruct the trend component. The threshold $\lambda = 0.0251$. The segmentation results are shown in Figure 22. The red dotted line is the result of spectrum segmentation based on the FEWT method. The number of boundaries is much less than with the EWT method; the calculation efficiency is greatly improved. The blue line is the kurtosis of each EM component. The spectrum of the maximum kurtosis is located at $[3869 \text{ Hz}, 5355 \text{ Hz}]$. 

![Figure 22](image)
The maximum EM component of kurtosis is taken out, as shown in Figure 23. Large noise can be seen in the waveform of EM8, but the periodic impact is not obvious. The envelope has a characteristic frequency of 102 Hz and 2–7 frequency doubling, and the fault characteristics are obvious. In summary, the bearing outer ring has a fault and the FEWT algorithm can classify fewer EM components than EWT when dealing with the fault of the outer ring of the rolling bearing, thus reducing the diagnosis time. Meanwhile, the maximum kurtosis component of EMs obtained by FEWT contains more characteristic frequencies and their frequency doubling. Therefore, the FEWT method can diagnose outer ring faults of rolling bearings and has obvious advantages over the EWT method.

5.2. Analysis of Bearing Inner Ring Fault Data

In the fast kurtogram proposed by Antoni [34], the spectrum kurtosis is calculated after decomposing the signal using the 1/3-binary tree filter bank. This section draws on the fast kurtogram to obtain a new kurtogram based on FEWT. Section 3.1 introduces the periodic impact when a rolling bearing fails and specifies the number of reconfiguration points as \( B \in [10, 60] \). It is necessary to study how to select the appropriate frequency component in the key function from this interval to reconstruct trend components and extract fault information. The fast kurtogram can accelerate the operation speed and show the position of periodic impulse signals in the frequency domain, and then get an estimation of center frequency and bandwidth. The kurtogram method based on FEWT inherits these advantages. The flowchart is shown in Figure 24.
Step 1: Since the number of reconstructed points affects the shape of the trend components, the number of the first reconstructed point is set in this section, named ResCot, and the initial value is 5.

Step 2: With ResCot as the boundary, the key function and trend component are obtained by FEWT. Finally, a series of signal components are obtained.

Step 3: Determine whether the number of reconstructed points exceeds the scope. If ResCot ∈ [5, 60], then set ResCot = ResCot + 5 and repeat step 2. If ResCot > 60, then stop cycling and calculate the kurtosis of all signal components and draw the kurtogram.

Step 4: Extract the components with the maximum kurtosis in the kurtogram; calculate envelope spectrum and extract faults.

In this experiment, the bearing fault test rig shown in Figure 25a is adopted to collect the vibration signal of the bearing inner ring fault data with type 6307. The motor speed is 1496 r/min, and the sampling frequency Fs = 15,360 Hz. By calculation, the rotation frequency of the bearing inner ring is 24.9 Hz and the fault characteristic frequency is 122.5 Hz. Figure 25b shows the signal waveform and Figure 25c shows its spectrum.

The experimental signal contains a large amount of noise. It is difficult to find the most suitable frequency band that may contain fault information from the spectrum. The signal processed by the FK
method can quickly find the frequency band with the maximum kurtosis and extract the information, as shown in Figure 26a. The eighth bands from the left side at level 4 with the maximum kurtosis are extracted. The noise is suppressed to a certain extent, and the periodicity of the impact is not obvious. Then the envelope is found, such as in Figure 26(a2). The characteristic frequency can be found from the envelope spectrum, but the high-frequency component cannot be found. The effect of extraction is not the most ideal.

![Figure 26. The inner ring: (a1) fast kurtogram based on the FIR filter; (a2) component in level 4 and its envelope; (b1) kurtogram based on FEWT; (b2) component with ResPoint = 24 and its envelope.](image)

Signal processing is done by the method mentioned in this paper. The kurtogram is shown in Figure 26b. When the number of reconstruction points ResPoint $\in [20; 36]$, although the number of reconstructed components is different, the frequency band of the component with the largest kurtosis is basically fixed. When the number of reconstructed points is 24, the maximum component of kurtosis appears. Its kurtosis is $K_{max} = 10.5$; the central frequency is $f_c = 3628$ Hz; the boundaries are $[3101.125$ Hz; $4155.125$ Hz]; the bandwidth is $Bw = 1054$ Hz. After extracting this component, it can be found that the component has an obvious periodic shock. There are obvious characteristic frequencies and frequency doubling in the envelope. By comparing the analysis results of FEWT and FK, the results obtained by FEWT are more reasonable. The extracted components have less noise and a more obvious impact. After obtaining the envelope, more frequency multiplication information can be obtained. The failure information is more intuitive. It can be seen that the fast empirical wavelet transform method is suitable for fault diagnosis of the bearing inner ring.

6. Conclusions

A new method named FEWT is proposed in this paper. The FEWT method can quickly decompose a signal into a series of empirical modes. The method determines the boundary through the minimum value of the trend component of the spectrum. This boundary is used to divide the spectrum and
reconstruct the components of the signal. The trend component can be obtained through the low-pass filtering of the Fourier transform function of the spectrum. The high number of reconstructed points reconstructs the complex trend, and the low number of reconstructed points reconstructs the rough trend. The simulation signal proves that the high number of reconstructed points and the threshold can successfully separate the two components, which are close to each other in the spectrum, but EWT cannot. The low number of reconstructed points can extract the components with side bands in the spectrum, which is more suitable for rolling bearing fault diagnosis.

The FEWT method improves two shortcomings of EWT: modal aliasing and invalid components. The success rate of separation proves that FEWT can separate effective components on the premise that it obtains less modality. The error rate shows that the modal separated by FEWT is more consistent with the original components.

Finally, the proposed method is proved to be better than the EWT method and fast kurtogram in extracting signal features by combining kurtosis.

Author Contributions: Y.X. and K.Z. conceived and designed the experiments. Y.X. and X.L. performed the experiments. Y.X. and K.Z. analyzed the data; C.M. and J.Z. provided guidance and recommendations for research; Y.X. contributed to the contents and writing of the manuscript. All authors have read and approved the final manuscript.

Funding: Funding was provided by the National Natural Science Foundation of China (Grant Nos. 51775005 and 51675009).

Acknowledgments: The authors would like to gratefully acknowledge the National Natural Science Foundation of China (grant nos. 51775005 and 51675009) and the Key Laboratory of Advanced Manufacturing Technology for their support. Finally, the authors would like to thank the editors and reviewers for their valuable comments and constructive suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References


