Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Sound and Vibration 286 (2005) 187–205
JOURNAL OF SOUND AND VIBRATION
www.elsevier.com/locate/jsvi

An improved Hilbert–Huang transform and its application in vibration signal analysis

Z.K. Peng[a], Peter W. Tse[b,*], F.L. Chu[a]

[a]*Department of Precision Instruments, Tsinghua University, Beijing 100084, P.R. China*
[b]*Smart Asset Laboratory, Department of Manufacturing Engineering and Management, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, P.R. China*

Received 10 September 2003; received in revised form 16 August 2004; accepted 1 October 2004
Available online 25 December 2004

Abstract

The vibration generated by industrial machines always contains nonlinear and non-stationary signals. Recently, a number of new methods have been proposed to analyse these signals. One of the promising methods is the Hilbert–Huang Transform (HHT). The HHT is derived from the principals of empirical mode decomposition (EMD) and the Hilbert Transform. When applying the HHT, first, the EMD will decompose the acquired signal into a collection of intrinsic mode functions (IMF). The IMF is a kind of complete, adaptive and almost orthogonal representation for the analysed signal. Since the IMF is almost monocomponent, it can determine all the instantaneous frequencies from the nonlinear or non-stationary signal. Second, the local energy of each instantaneous frequency can be derived through the Hilbert Transform. Hence, the result is an energy–frequency–time distribution of the signal. Since applying the process of HHT is not computational intensive, the HHT becomes a promising method to extract the properties of nonlinear and non-stationary signal. However, after the completion of a thorough experiment, the result generated by the HHT has its deficiency. First, the EMD will generate undesirable IMFs at the low-frequency region that may cause misinterpretation to the result. Second, depends on the analysed signal, the first obtained IMF may cover too wide a frequency range such that the property of monocomponent cannot be achieved. Third, the EMD operation cannot separate signals that contain low-energy components. In this study, new techniques have been applied to improve the result of HHT. In the improved version of HHT, the wavelet packet transform (WPT) is used as preprocessing to decompose the signal into a set of narrow band signals prior to the application of EMD. With the help from WPT, each

*Corresponding author.
E-mail addresses: pengzhike@tsinghua.org.cn (Z.K. Peng), meptse@cityu.edu.hk (P.W. Tse).

0022-460X/$ - see front matter
doi:10.1016/j.jsv.2004.10.005

IMF derived from the EMD can truly become monocomponent. Then, a screening process is conducted to remove unrelated IMFs from the result. Both simulated and experimental vibration signals of having a rotary system with the fault of rubbing occurred have proven that the improved HHT does show the rubbing symptoms more clear and accurate than the original HHT. Hence, the improved HHT is a precise method for nonlinear and non-stationary signal analysis.

1. Introduction

Signal analysis has always been an important and indispensable part in vibration-based machine fault diagnosis as well as in many other practical applications. In the application of vibration-based machine fault diagnosis, the signal analysis often serves two purposes: to investigate the dynamical characters of machines under different fault types, and to extract fault features if a fault occurs and then identify the cause of fault. Hitherto, the technique of Fourier Transform has been dominated in the field of signal analysis because of its prowess and simplicity. However, there are some crucial restrictions on the use of Fourier transform [1]. The signal generated by the inspected machine must be linear and temporally stationary; otherwise, the resulting Fourier spectrum will have little physical sense. Unfortunately, in vibration-based machines fault diagnosis, the signals to be analysed are often non-stationary and nonlinear. The resulting frequency components are not always consistent as the patterns of the acquired signals often change with time. Therefore, Fourier Transform cannot fulfil the requirements of fault diagnosis, particularly in real applications.

From the past decade, wavelet transform has become one of the fast-evolving mathematical and signal processing tools [2]. Wavelet transform is complete, orthogonal (in the discrete form), local and adaptive. All these are vital for forming a basis to analyse nonlinear and non-stationary signals. The basic operation of wavelet transform involves the operations of dilation and translation, which lead to a multiscale analysis of the signal. Hence, it can extract both the time and frequency features of the inspected signal effectively. Although wavelet transform is capable of analysing nonlinear and non-stationary signals and deemed suitable for vibration-based machine fault diagnosis, many deficiencies have been reported in the use of wavelet transform [3]. The inevitable deficiencies include the interference terms, border distortion and energy leakage. These deficiencies may generate a lot of undesired small spikes all over the frequency scales and make the results confusing and difficult to be interpreted. Basically, wavelet transform can be categorized into continuous wavelet transform and discrete wavelet transform. The process of continuous wavelet transform is computationally intensive. It is very time-consuming in analysis if the acquired data are numerous. On the other hand, the discrete wavelet transform has good computing efficiency. However, the resolution in frequency at a high-frequency range is poor. Hence, it may be applicable to many applications, such as data compression and noise removal, but not for frequency analysis in high-frequency range.

Due to the deficiencies of wavelet transform, a new type of time–frequency analysis called Hilbert–Huang transform (HHT) has been proposed for analysing nonlinear and non-stationary

signals [4,5]. The HHT is derived from the principals of empirical mode decomposition (EMD) and the Hilbert Transform. When applying the HHT, first, the EMD will decompose the acquired signal into a collection of intrinsic mode functions (IMF). The IMF is a kind of complete, adaptive and almost orthogonal representation for the analysed signal. Since the IMF is almost monocomponent, it can determine all the instantaneous frequencies from the nonlinear or non-stationary signal. Second, the local energy of each instantaneous frequency can be derived through the Hilbert Transform. Hence, the result is a HHT spectrum which has an energy–frequency–time distribution of the signal. From the HHT spectrum, one can localize any event on its occurring time as well as its instantaneous frequency. Comparing wavelet transform to the EMD, which is the most computational intensive process in the HHT, the EMD does not involve any convolution. Hence, for EMD, the time used for computation is less and deemed suitable for analysing numerous data or signals. The technique of HHT has been used for vibration signal analysis in a number of applications. Yang and Sun [6] used the HHT to interpret the nonlinear response of a crack-induced rotor. Yang and Lei [7] proposed an HHT-based damage identification approach and applied it to the ASCE structural health monitoring benchmark structure.

Although the HHT may become a promising method to extract the properties of nonlinear and non-stationary signal, like other signal analyses, HHT also suffers from a number of shortcomings. First, the EMD will generate undesirable IMFs at the low-frequency region that may cause misinterpretation to the result. Second, it depends on the analysed signal, the first obtained IMF may cover too wide a frequency range that the property of monocomponent cannot be achieved. Third, the EMD operation cannot separate signals that contain low-energy components. To solve these shortcomings, an improved HHT is presented here. The improved HHT uses the wavelet packet transform (WPT) [8] as a preprocessor to separate the inspected signal into a set of narrow band signals. Hence, frequency components that contain low energy are easier to be identified at different narrow bands. Then the process of EMD will be applied to decompose these narrow band signals as the original EMD will do. The resultant IMFs will be in narrower bands and become easier to satisfy the condition of monocomponent. By applying both WPT and the EMD, the second and the third shortcomings aforementioned can be avoided. Then, a screening process will be conducted to select the vital IMFs from the unrelated IMFs. This can be achieved by calculating the correlation coefficients of the IMFs with the inspected raw signal. Based on the values of the coefficients, unrelated IMFs that may cause distortion to the results, particularly in low-frequency range as mentioned in the first shortcoming, can be minimized. Simulated signals of rubbing occurring between a rotor and a stator, and real signals generated from a rotary machine that has the fault of rubbing, have been used to verify the effectiveness of the improved HHT.

The definition of instantaneous frequency, the deficiency of Hilbert transform, and the principal of HHT, which consists of the EMD and the Hilbert transform, are introduced in Section 2. Section 3 discussed the rationale to develop the improved HHT, the theory of the improved HHT, and its formation. The construct of the simulated data, the experimental set-up for generating faulty vibration from a real rotary machine that has the symptom of rubbing, and the comparison between the improved HHT and the original HHT in the capability of detecting rubbing, are presented in Section 4. The conclusion and the potential of the improved HHT are stated in Section 5.

2. Instantaneous frequency and Hilbert–Huang transform

2.1. Definition of instantaneous frequency and capability of Hilbert transform

Although the definition of instantaneous frequency is always controversial, it is tenable to define that for a given length of signal, there is only one frequency value within the length of the signal, or the signal is monocomponent. To extract the instantaneous frequency of a monocomponent signal, the Hilbert transform can be used. For an arbitrary signal $x(t)$, its Hilbert transform $y(t)$ is defined as

$$y(t) = \frac{P}{\pi} \int_{-\infty}^{+\infty} \frac{x(\tau)}{t-\tau} \, \mathrm{d}\tau, \tag{1}$$

where P is the Cauchy principal value. From Eq. (1), it can be seen that the Hilbert transform is defined as the convolution of the signal $x(t)$ with $1/t$ [1]. Therefore, the Hilbert transform is capable of identifying the local properties of $x(t)$. Coupling the $x(t)$ and $y(t)$, we can have the analytic signal $z(t)$ of $x(t)$, as

$$z(t) = x(t) + \mathrm{i}y(t) = a(t)\mathrm{e}^{\mathrm{i}\varphi(t)}, \tag{2}$$

where

$$a(t) = [x^2(t) + y^2(t)]^{1/2}, \quad \varphi(t) = \arctan(y(t)/x(t)). \tag{3}$$

The $a(t)$ is the instantaneous amplitude of $x(t)$, which can reflect how the energy of the $x(t)$ varies with time, and the $\varphi(t)$ is the instantaneous phase of $x(t)$. The controversial instantaneous frequency $\omega(t)$ is defined as the time derivative of the instantaneous phase $\varphi(t)$, as follows:

$$\omega(t) = \frac{\mathrm{d}\varphi(t)}{\mathrm{d}t}. \tag{4}$$

The temporal waveform of a chirp signal, which is a linear frequency modulation signal, and its instantaneous frequency calculated by Eq. (4) are shown in the top and bottom diagram of Fig. 1, respectively. The scale of x-axis in both diagrams is in a number of sample points. The scale of y-axis in the instantaneous frequency spectrum is in normalized frequency. Similar orientation and scaling, that is, the sample point represents the scale in time and the normalized frequency represents the scale in frequency, will be applied to the rest of the figures in this paper. From the instantaneous frequency spectrum, the frequency of the chirp signal changes linearly from 0 to 0.5 in its lifespan. That is, the calculated instantaneous frequencies can truly reveal the frequency properties of the chirp signal. Therefore, Eq. (4) is useful in extracting instantaneous frequencies from the non-stationary chirp signal. However, Eq. (4) is only valid in obtaining the instantaneous frequency of a signal in a given time frame if the signal is monocomponent within the time frame. If the inspected signal is multicomponent within the defined time frame signals, the result of the instantaneous frequency will be distorted.

The top diagram of Fig. 2 shows a multicomponent signal, which contains two frequency components. Its instantaneous frequency has been calculated by Eq. (4) and displayed in the bottom diagram of Fig. 2. As shown in the bottom diagram, the calculated instantaneous frequency ranges from 0.25 to 0.5, instead of only two frequency components. Hence, Eq. (4) fails

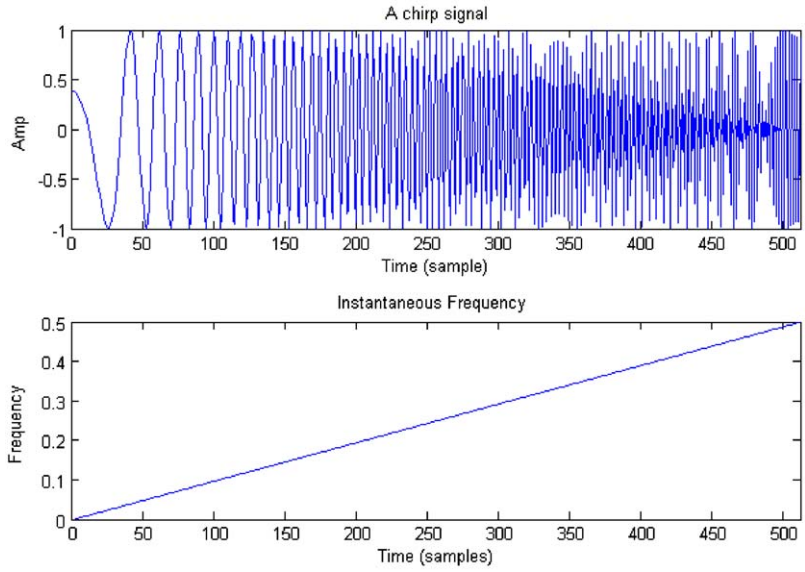

Fig. 1. A chirp signal and its calculated instantaneous frequency.

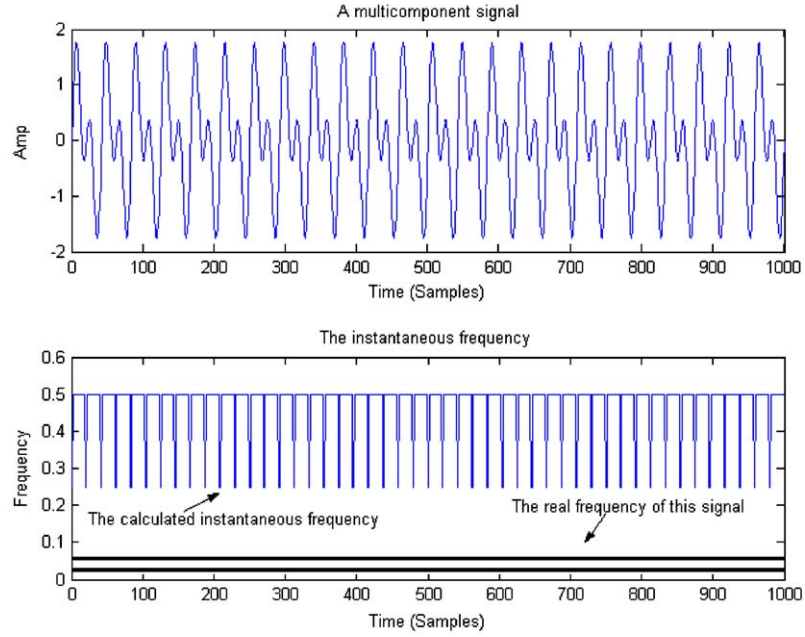

Fig. 2. A multicomponent signal and its calculated instantaneous frequency.

to reflect the instantaneous frequency of the multicomponent signal. Unfortunately, in almost all of the practical applications, the inspected signals are hardly monocomponent but multicomponent. The signals usually have more than one instantaneous frequency at a time, and the

Hilbert transform fails to reveal the true frequency content of the inspected signals. Therefore, to make the instantaneous frequency applicable, the key is the ability to decompose a signal into some individual monocomponent signals to which the basic definition of the instantaneous frequency can be applied. The EMD provides such decomposition ability to separate multicomponent signal into some almost orthogonal, nearly monocomponent signals, or in other terms, into different IMF components. Then, the Hilbert transform can be applied to identify the instantaneous frequencies. Therefore, the HHT, which utilizes the EMD as a preprocessor and then obtain the instantaneous frequency by the Hilbert transform, can be applicable to any nonlinear and non-stationary signal.

2.2. Empirical mode decomposition and intrinsic mode function

Huang et al. [1] presented the use of EMD to decompose any multicomponent signal into a set of nearly monocomponent signals and are referred as IMFs. Once the IMFs are obtained, then the instantaneous frequency of each IMF can be determined. Physically, the necessary conditions to define a meaningful instantaneous frequency are that the inspected signal must be symmetric with respect to the local zero mean, and have the same numbers of zero crossings and extrema. That is, in an IMF function, the number of extrema and the number of zero crossings must either equal or differ at most by one in whole data set, and the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero at every point. However, the determined IMF may not satisfy these conditions precisely. Therefore, the resulting IMF is nearly a monocomponent signal, but not perfectly.

The EMD is developed based on the assumption that any signal consists of many different IMFs. The procedures to decompose a given signal $x(t)$ to different IMFs can be categorized into the following steps. First, identify all the local extrema from the given signal, and then connect all the local extrema with a cubic spline line as the upper envelope. Second, repeat the first step for the local minima to produce the lower envelope. The upper and lower envelopes should cover the entire signal between them. Third, designate their mean as m_1, and the difference between the signal $x(t)$ and m_1 as the first component h_1, that is,

$$x(t) - m_1 = h_1. \tag{5}$$

Ideally, after the sifting operation of Eq. (5), h_1 should be an IMF. The construction of h_1 described above seems to have satisfied all the requirements of IMF. However, during the process, overshoots and undershoots may exist and could be classified as new extrema. A few overshoots and undershoots may shift or exaggerate the existing ones, ultimately distorting the means. Moreover, the envelope mean may be different from the true local mean for nonlinear signal, which may make h_1 asymmetric. To eliminate riding waves and make the wave profiles more symmetric, Huang et al. repeated the sifting process of Eq. (5) as many times as required to reduce the extracted signal to an IMF. Therefore, the fourth step is to repeat the sifting process by treating h_1 as the signal and repeat Eq. (5) as

$$h_1 - m_{11} = h_{11}. \tag{6}$$

The sifting process will be repeated k times, until h_{1k} becomes a true IMF, that is,

$$h_{1(k-1)} - m_{1(k-1)} = h_{1k}, \tag{7}$$

then it is designated as

$$c_1 = h_{1k}. \tag{8}$$

Finally, we obtained the first IMF component from the signal.

Huang et al. also suggested a criterion for stopping the sifting process. This is accomplished by limiting the size of the standard deviation, denoted as S.D., which is calculated from two consecutive sifting results as

$$\text{S.D.} = \sum_{t=0}^{N} \left[\frac{\left| h_{1(k-1)}(t) - h_{1k}(t) \right|^2}{h_{1(k-1)}^2(t)} \right]. \tag{9}$$

According to Huang et al., the S.D. value of 0.2–0.3 for the sifting process is a very rigorous limit for the difference between two consecutive siftings.

Generally, c_1 should contain a component that has the finest scale or the shortest period of the signal. Removing c_1 from the rest of the signal by

$$x(t) - c_1 = r_1 \tag{10}$$

then we will have the residue of the signal r_1, which contains a component with a longer period than the previous component. Treating r_1 as a new signal and repeating the same sifting process as described above, we can then obtain the second IMF c_2. Similarly, we can obtain a series of IMFs $c_i (i = 1, 2, \ldots, n)$ and the final residue r_n. The sifting process can be stopped by any of the following predetermined criteria: either when the component c_n or the residue r_n becomes less than the predetermined value of substantial consequence, or when the residue r_n becomes a monotonic function from which no more IMF can be extracted. Summing up all the IMFs and the final residue r_n, we should be able to reconstruct the original signal $x(t)$ by

$$x(t) = \sum_{i=1}^{n} c_i + r_n. \tag{11}$$

After obtaining all the IMFs, the Hilbert Transform can be applied to each IMF and calculate the instantaneous frequency according to Eqs. (3) and (4). Now the original signal $x(t)$ can be expressed as

$$x(t) = \sum_{j=1}^{n} a_j(t) \exp\left(\mathrm{i} \int \omega_j(t)\, \mathrm{d}t \right). \tag{12}$$

Eq. (12) enables us to represent the inspected signal in its instantaneous amplitude, frequency, and time in a three-dimensional plot or a contour map. The distribution of the signal's amplitude in a time–frequency plot is designated as the Hilbert spectrum, $H(\omega, t)$.

3. The principal of the improved Hilbert–Huang transform

The fundamental theory of the improved HHT has been presented in an authors' paper which presents a comparison study between the improved HHT and wavelet transform in roller bearing fault diagnosis [9]. Nevertheless, the detailed theory and formation of the improved HHT are

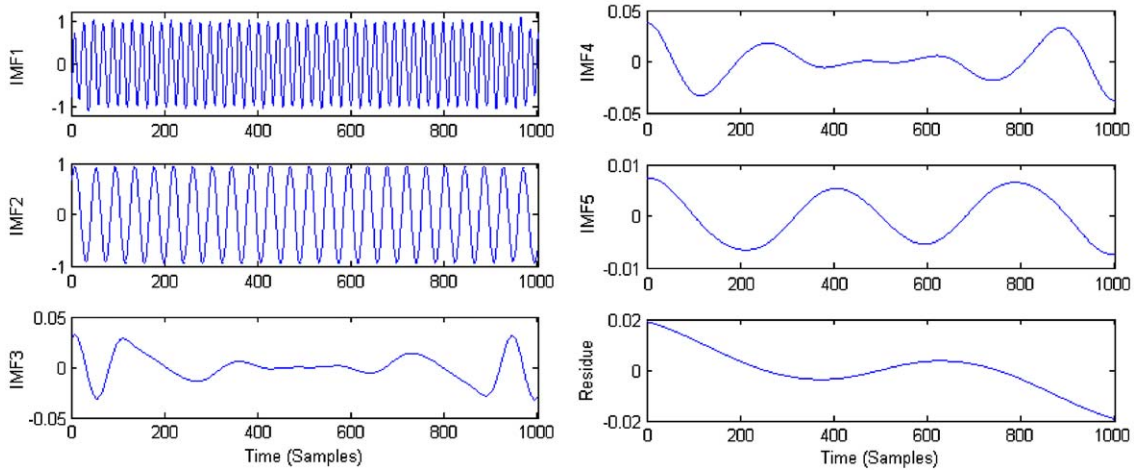

Fig. 3. The IMFs and the residue of the multicomponent signal described in Fig. 2.

described in this section. As mentioned in Section 2, the HHT could be a promising tool for analysing non-stationary and nonlinear signals. However, after conducting a comprehensive study, shortcomings of the HHT have been found. The three major shortcomings include the EMD generating undesirable IMFs at the low-frequency region, the first obtained IMF may cover too wide a frequency range such that the property of monocomponent cannot be achieved, and some signals that contain low-energy components are inseparable. Fig. 3 displays the results after applying the original HHT to the signal as shown in the top diagram of Fig. 2. The signal has been decomposed into five IMFs and its residues. Note that prior to decomposition, the inspected signal only contained two frequency components. Therefore, the first two IMFs (IMF1 and IMF2) are real components of the signal, whilst, the others are pseudo-components in low-frequency range (IMF3, IMF4, and IMF5). Such pseudo-components may mislead inexperienced machine operators in using the decomposed signals for vibration-based fault diagnosis. These undesirable IMFs can be removed by applying a screening process, which will retain the essential IMFs that are relevant to the inspected signal.

Since the combined IMFs and its residue should be an orthogonal representation of the inspected signal, the relevant IMFs should have strong correlation with the signal, whilst, the irrelevant IMFs should posses weak correlation with the signal. Based on the above argument, the correlation coefficient μ for each IMF with the inspected signal will be calculated and acted as a criterion for selecting relevant IMFs. The calculated correlation coefficients will be normalized to avoid accidental removing of some low-amplitude but relevant IMFs. A threshold can be set for the screening process. For all correlation coefficients μ_i where $i=1,\ldots n$; n is the number of IMF), the threshold λ is set by the ratio of the maximal μ_i, which is

$$\lambda = \max(\mu_i)/\eta, \quad (i = 1, \ldots, n), \tag{13}$$

where η is a ratio factor. In this study, $\eta = 10.0$ is used. Any IMF that has a correlation coefficient equal or higher than λ will be retained, whilst, other IMFs will be removed and added to the residue r_n.

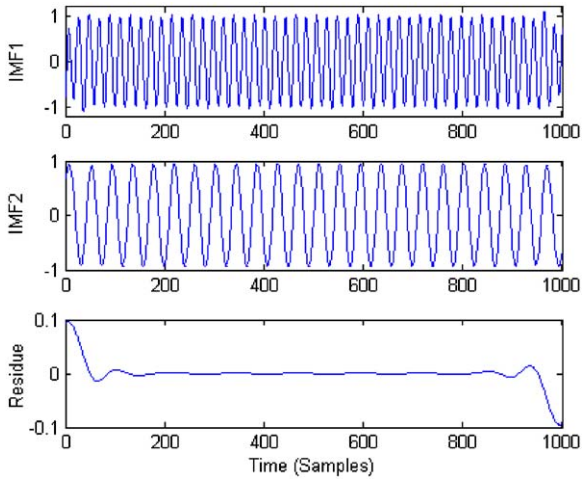

Fig. 4. The result after applying the screening method for selecting relevant IMFs.

For the signal that contains only two frequency components, the first two IMFs have a correlation coefficient above 0.7 while the other three IMFs have a correlation coefficient below 0.05. Hence, the screening method will select the first two IMFs, which are IMF1 and IMF2, and add the other remaining IMFs to the residue. The result of the first two IMFs and the residue is shown in Fig. 4. Note that the new residue, which has significant swings at both ends of the residue, is substantially different than the one shown in Fig. 3. It is mainly because the irrelevant IMFs have been added to the residue.

As refers to the bottom diagram of Fig. 2, the original HHT fails to identify the instantaneous frequency in the Hilbert spectrum. After applying the screening method for selecting relevant IMFs, the improved Hilbert spectrum can now display the two frequency components as expected from the inspected signal as shown in Fig. 5. Even the change of frequency in time can be highlighted on the Hilbert spectrum. Although the appearance of the Hilbert spectrum is similar to the scalogram generated by wavelet transform, the Hilbert spectrum does not involve the dilemma of fine frequency resolution but poor time resolution, or fine time resolution but poor frequency resolution as often occurs in wavelet scalogram. The Hilbert spectrum will only reveal the real pattern of instantaneous frequency for the inspected signal.

The second shortcoming of the original HHT is the first generated IMF may cover too wide a frequency range that the property of monocomponent cannot be achieved. The top left-hand and right-hand-side diagrams of Fig. 6 show the temporal waveform of a simulated faulty signal caused by rubbing and its FFT spectrum, respectively. The decomposed IMF1 to IMF3 and their FFT spectra are shown in the successive left-hand- and right-hand-side diagrams, respectively. Note that the temporal waveform of the first IMF (second row left-hand-side diagram) contains more than one frequency component. Such phenomenon reflects in its FFT spectrum (second row right-hand-side diagram) by having a number of FFT lines. Obviously, the IMF1 fails to satisfy the property of monocomponent. Nevertheless, the waveforms of IMF2 and IMF3 and their FFT spectra displayed in the successive diagrams show that they are nearly monocomponent.

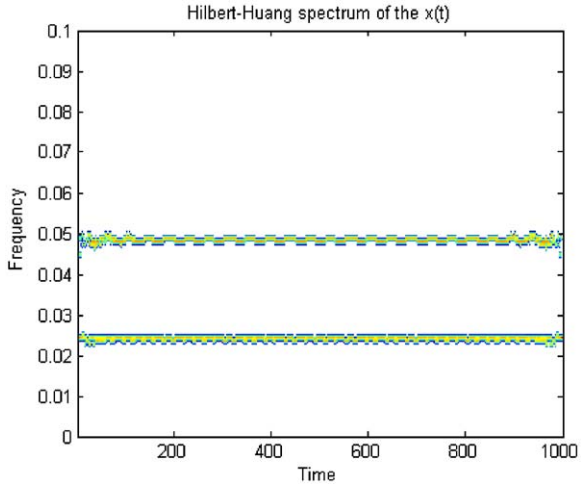

Fig. 5. The improved Hilbert spectrum can clearly show the two frequency components of the inspected signal and their changes in time.

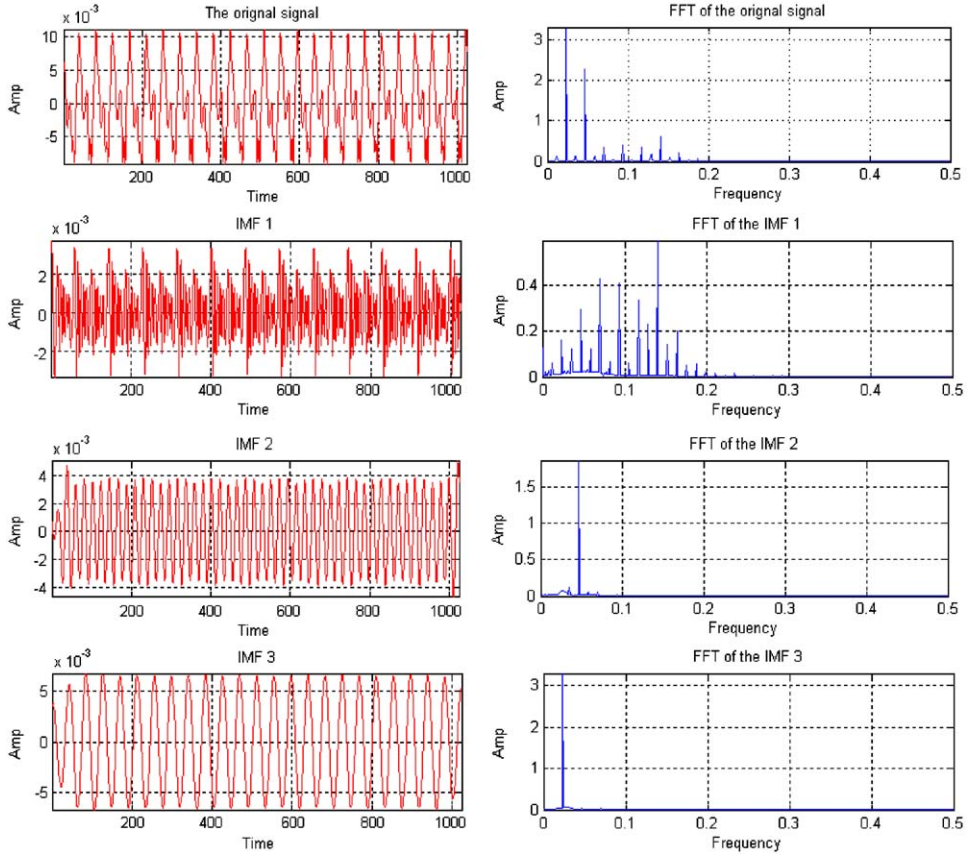

Fig. 6. A simulated rubbing signal demonstrated the first IMF fails to maintain the property of monocomponent.

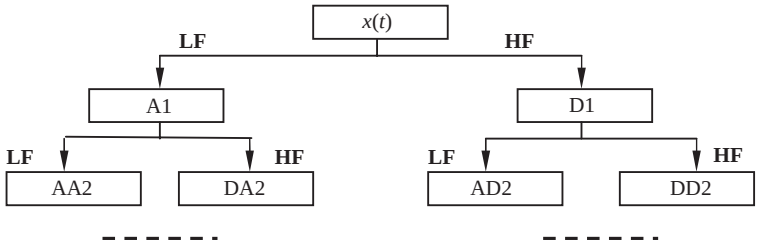

Fig. 7. The WPT decomposition tree.

The third shortcoming is during the EMD operation, some low-energy components may be missing. As shown in the FFT spectrum of the IMF1 (second row right-hand-side diagram of Fig. 6), some low-energy components do exist. When the simulated signal is required to be reconstructed by its decomposed IMFs and residue, the low-energy components may be masked by other high-energy components (some of them are shown in the IMF2 and IMF3 diagrams). Hence, these low-energy components will not be shown as expected in the frequency–time plane of the reconstructed signal.

From the property of monocomponent, each decomposed IMF should represent a simple oscillatory mode that is embedded in the inspected signal. Within each cycle, defined by the zero crossing, the decomposed IMF should involve only one mode of oscillation without complex riding waves. According to the above definition, each IMF is not restricted to a narrow band signal. It can be a signal that has been both amplitude and frequency modulated. To ensure the property of monocomponent is valid for each decomposed IMF, a suggestion is to decompose the signal into some narrow band signals first. Then the operation of EMD will be applied to each narrow band signal. Hence, the obtained IMFs will be in narrow frequency bands and their instantaneous frequencies can be calculated by Eq. (4).

There are a number of methods that can be used to act as a preprocessor to separate the inspected signal into various narrow band signals. Among them, the WPT can be a proper preprocessor due to its well-known properties of being orthogonal, complete, and local. As illustrated in Fig. 7, during the operation of WPT, an inspected signal, $x(t)$, will be split into an approximation portion (A1) and a detail portion (D1) through a couple of low-band filter (LF) and high-band filter (HF), respectively. The approximation portion (A1) will then further split into a second-level approximation portion (AA2) and a detail portion (DA2) by the LF and HF, respectively. Similarly, the first-level detail portion (D1) will also split into a second-level approximation portion (AD2) and a detail portion (DD2). The process will be continued until a stopping criterion has been reached. For an n-level decomposition, the signal will be decomposed into 2^n narrow band signals. Note, indifference to the energy content of the component, each component will be decomposed into its bands. That is, even a component has low energy content, unlike the EMD operation, the WPT process will retain its property.

The HHT with the WPT as preprocessor plus the IMFs selection method is here called the improved HHT. Its effectiveness will be verified in the coming section by using both signals generated from a simulated rubbing fault and a real rotary system with a rubbing fault occurring in the system. The operation procedures of the improved HHT can be summarized in the flowchart as shown in Fig. 8.

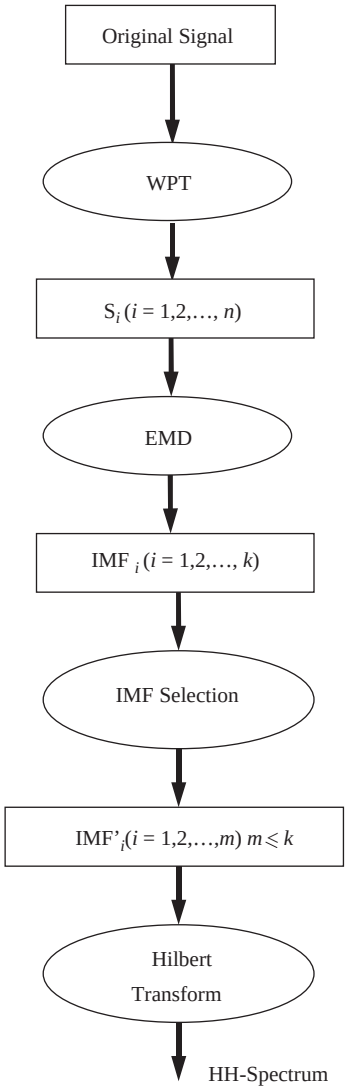

Fig. 8. The operation procedures of the improved HHT.

4. The use of improved HHT in rubbing vibration signal analysis

For any rotary machine, rubbing often occurs between the surface of a rotor and a stator. The symptom of rubbing occurs at the position if the clearance between a rotor and a stator is too tight. If the symptom of rubbing has been ignored by the machine operator, the damage of rubbing will be amplified and finally lead to a catastrophic breakdown of the rotary machine. For example, the continuous rubbing between a motor's rotor and its stator could cause fatal breakdown of a motor. Rubbing between a blade and its seals could result in a broken seal. The dynamic behaviour of the vibration signal generated by rubbing is often nonlinear and complicated [10]. In order to minimize the damage caused by rubbing, a number of methods have

been proposed to detect rubbing at its early stage of occurrence [11–13]. In order to compare the effectiveness of the original HHT and the improved HHT, vibration signals generated by a simulated rubbing and a real rotary system that has the fault of rubbing, have been used for verification.

4.1. Comparison in analysing simulated rubbing signals

Three sets of rubbing vibration signals, which include slight, moderate, and serious rubbing conditions, were generated from a simulated single-disk rub-impact rotor model [14]. Fig. 9 shows the temporal signal (bottom left diagram) and its FFT spectrum (bottom right diagram) simulated under slight rubbing condition. The scale of x-axis is the time that the number of sample points required to be collected. The scale of y-axis is the nominated amplitude of the inspected signal. The results generated by applying the original HHT and the improved HHT to the signal are shown in the top left and top right diagrams, respectively. The scale of x-axis is the same as before, and the scale of y-axis is the nominated frequency of the inspected signal. The contours shown in the diagrams represent the amplitude of the results. The greyer the colour, the higher the amplitude of the result. Hereinafter, similar labels, notations, and scales will be used for all the figures and their diagrams.

As mentioned in Section 3, the original HHT will produce the first IMF (IMF_1) that will cover too wide a frequency range and fails to be treated as a monocomponent signal. Due to this deficiency, the instantaneous frequency calculated by Eq. (4) cannot fully represent the real

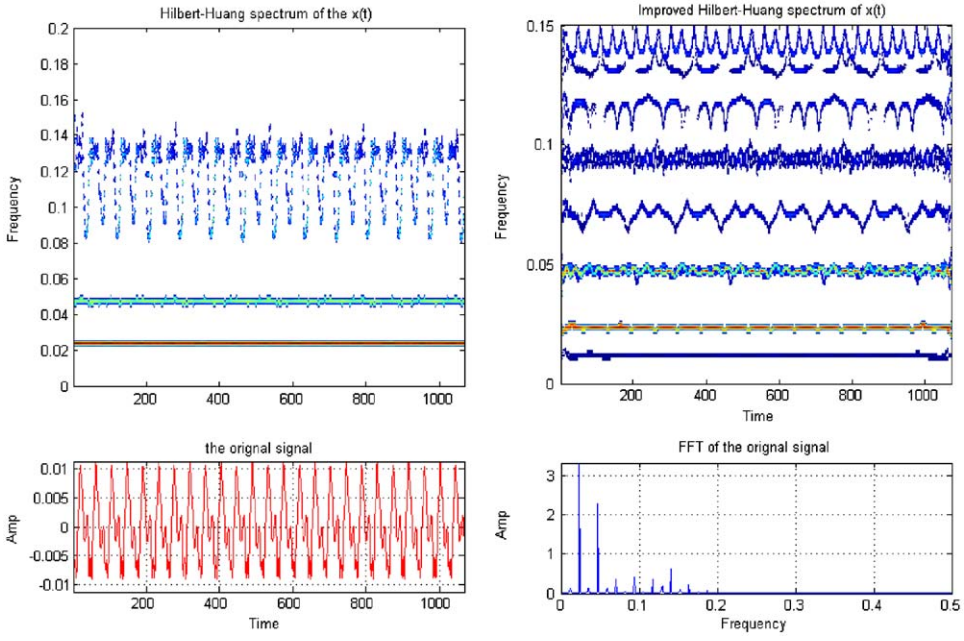

Fig. 9. The temporal waveform of a simulated vibration signal with slight rubbing condition, its FFT spectrum, and the results of the original HHT and the improved HHT.

frequency pattern of the IMF_1. Sequentially, the Hilbert spectrum obtained by the original HHT fails to reflect the true frequency pattern of the inspected signal. In the result of the original HHT as shown in the top left diagram of Fig. 9, the frequency patterns from the frequency range of 0.07–0.15 are hardly being identified. The results only show that the change of frequency is periodic; similarly the temporal signal as shown in the original raw signal in the bottom left diagram. However, the detailed change of each instantaneous frequency is difficult to be revealed as it cannot be separated from others. On the other hand, from the results generated by the improved HHT spectrum as shown in the top right diagram of Fig. 9, the detailed change of each instantaneous frequency can be clearly revealed. Starting from the frequency range of 0.07–0.15, the high-frequency portion of the inspected signal, each frequency component embedded in the signal and its instantaneous changes are well separated from other component. Hence, the operators can easily identify the patterns of the signal both in time and frequency. Such ability will definitely help the operators in determining any anomaly occurring in the signal.

Another important discovery here is that the improved HHT can recover frequency component that contains low energy. As mentioned in Section 3, the original HHT has difficulty in recovering such frequency component. The improved HHT can enhance this ability. As shown in the top right diagram of Fig. 9, a low-frequency and low-energy component at round 0.01, which has been missed in the results generated by the original HHT, has been recovered by the improved HHT. In the rubbing fault diagnosis, low-frequency component may contribute the sign of a rotary system starting to deteriorate due to rubbings. From another study conducted by the authors, when rubbings occur in a rotary system, two symptoms can be observed [11]. First, the low-frequency components of the rubbing signal often have relative constant frequency values in time. Second, the high-frequency components will have fluctuated values that is often changing periodically in time. Such symptoms can be revealed from the results of the improved HHT as shown in the top right diagram of Fig. 9. The high-frequency components from 0.07 to 0.15 are changing periodically with time, whilst, the low-frequency components from 0.01 to 0.05 are not having much change in time. Hence, even a low-frequency component that has low energy content, it cannot be ignored, as it may contribute an early warning of the occurrence of rubbings.

The temporal waveform of the vibration signal simulated by a moderate rubbing condition, its analysed FFT spectrum, and the comparison results of different HHTs are shown in Fig. 10. Similar remarks observed from Fig. 9 can also be found in Fig. 10. Both details of the instantaneous frequencies in the high-frequency range and components that have low energy content in the low-frequency range can be revealed in the results of the improved HHT as shown in the top right diagram of Fig. 10. Again, the improved HHT can present the true frequency patterns of the instantaneous frequencies. This time, the low-energy frequency component around 0.01 also appears in the original HHT's results. Actually, this component around 0.01 is a fractional harmonic of the fundamental frequency of the rotary system. It is often used to evaluate the severity in rubbings. As rubbings continue to increase, the harmonic becomes more obvious. Hence, it appears in the original HHT's results as the rubbing condition has deteriorated to moderate. Note that in the improved HHT's results, the high-frequency components from 0.07 to 0.17 change their frequency values periodically, particularly for the components near 0.07, 0.12 and 0.14, whilst, the low-frequency components change their amplitudes periodically by showing

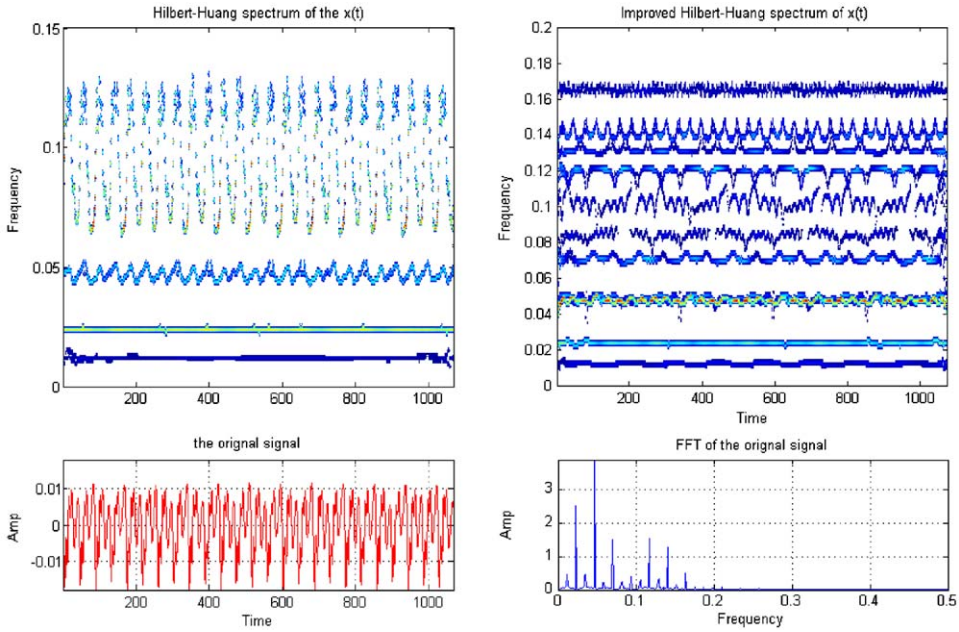

Fig. 10. The temporal waveform of a simulated vibration signal with moderate rubbing condition, its FFT spectrum, and the results of the original HHT and the improved HHT.

different grey scale of amplitude in time. An obvious one is the low-frequency component at around 0.05.

Fig. 11 shows the waveform of the inspected signal and the results for a simulated signal under serious rubbing condition. Compared to the analysed results for slight and moderate rubbing conditions, the analysed results for serious rubbing condition have more frequency components that are closely packed to each other. Again, the results generated by the original HHT (top left diagram) fail to present the details of instantaneous frequencies, particularly, in the high-frequency range. On the other hand, the results of the improved HHT reveal all the details clearly. Even the high-frequency components that are closely packed to each other can also be separated by the improved HHT, such as the components near 0.14.

A distinct observation from the original HHT's results is the low-frequency components also show periodic changes for the components located within the frequency range of 0.03–0.05. Such a phenomenon is not supposed to have happened for low-frequency component in rubbing. On the other hand, in the results of the improved HHT (top right diagram), in the frequency range from 0.04 to 0.06, there are two frequency components instead of one component as shown in the original HHT's results. Hence, the original HHT has distorted the true results. Such distortion will definitely affect the accuracy in vibration signal analysis, which may lead to an incorrect conclusion in fault diagnosis. Note that, when the rubbing condition becomes serious, more notable high-frequency components change themselves with time. Such an observation is an accordance with the two symptoms previously defined for monitoring the deterioration of rubbing.

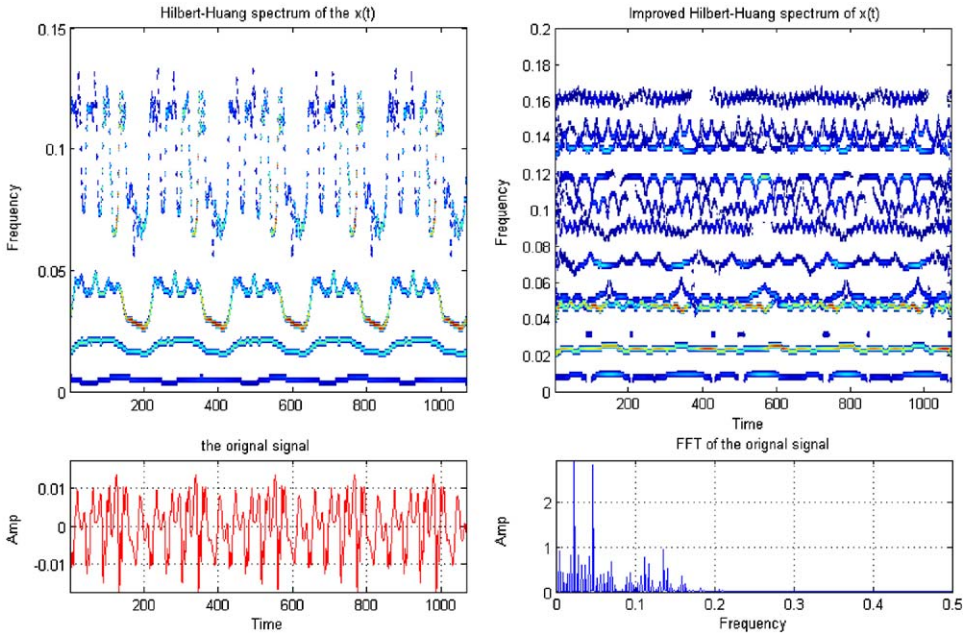

Fig. 11. The temporal waveform of a simulated vibration signal with serious rubbing condition, its FFT spectrum, and the results of the original HHT and the improved HHT.

4.2. Comparison in detecting rubbings occurred in a real rotary system

To further verify the performance of the improved HHT, vibration signals were collected from a real rotary system that had a moderate and later a serious rubbing condition. The rotary system consists of a rotor and a stator, a driving motor, journal bearings and couplings. Different conditions of rubbings were created by increasing the tightness between the rotor and the stator so that both the surfaces and the severity of rubbings increased. More detailed information on the rotary system and the construction of rubbings can be found in Ref. [11]. Vibration signals were collected from the rotary system using non-contact eddy-current transducers at a sampling rate of 1.6 kHz. The rotational speed of the rotary system was set at 3000 rev/min.

Fig. 12 shows the temporal waveform of the vibration signal acquired from the rotary system suffered from a moderate rubbing condition, its analysed and the FFT spectrum, and the comparison results of the original HHT and the improved HHT. The labels, notations, and scales used for previous figures also apply here. Similar to previous comparison results, except for the frequency component near 0.05, the results generated by the original HHT (the top left diagram) cannot present the true frequency patterns of the inspected signal collected under the moderate rubbing condition. Again, the high-frequency components are so vague that they are hardly being identified. The frequency component near 0.025 is missing because of its low energy content. On the contrary, the results of the improved HHT (the top right diagram) show more frequency components and separate the high-frequency components clearly. The missing low-frequency component near 0.025 can easily be identified.

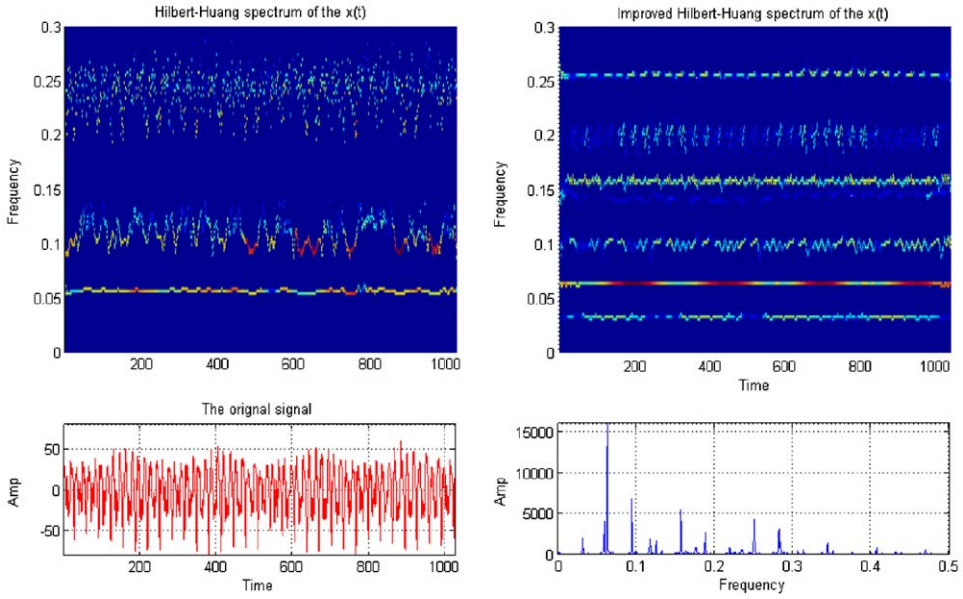

Fig. 12. The temporal waveform of a real vibration signal collected from moderate rubbings, its FFT spectrum, and the results of the original HHT and the improved HHT.

The two symptoms of rubbing also become obvious in the results of the improved HHT. The high-frequency components fall within the range of 0.15–0.3 change their frequency values periodically, whilst, the low-frequency components adjacent to 0.05 change their amplitudes periodically, but hold their frequency values relatively constant in time. Therefore, the two symptoms for determining rubbings occurred to appear in both simulated and real experimental tests.

The temporal waveform of the vibration signal collected from serious rubbings occurring in the rotary system, the signal's FFT spectrum, and the results of the original HHT and the improved HHT, are shown in Fig. 13. Compared with the signal acquired during the rotary system suffered from moderate rubbings, the results of the improved HHT (the top left diagram) show more high-frequency components, such as the one at around 0.27. Similar to the simulated results, the low-frequency component near 0.025, which has been missing in the results of the original HHT as in Fig. 12, appears in the results of the original HHT because of the increased severity in rubbings. Nonetheless, the high-frequency components are still very difficult to be identified in the results of the original HHT (the top right diagram). Again, the results of the improved HHT present the true frequency patterns of the inspected signal. The two major symptoms for confirming the existence of rubbings also appeared in the results of the improved HHT, with both the changes of frequency for high-frequency components locate within 0.18–0.3 and the changes of amplitude for the low-frequency components near 0.05 appear nearly periodically with time.

Both the experimental and the simulated results confirm that the improved HHT can provide a clear diagnosis on the symptoms of rubbings. Hence, it is a promising signal analysing for detecting the occurrence of rubbing fault and monitoring its deterioration. On the contrary, the

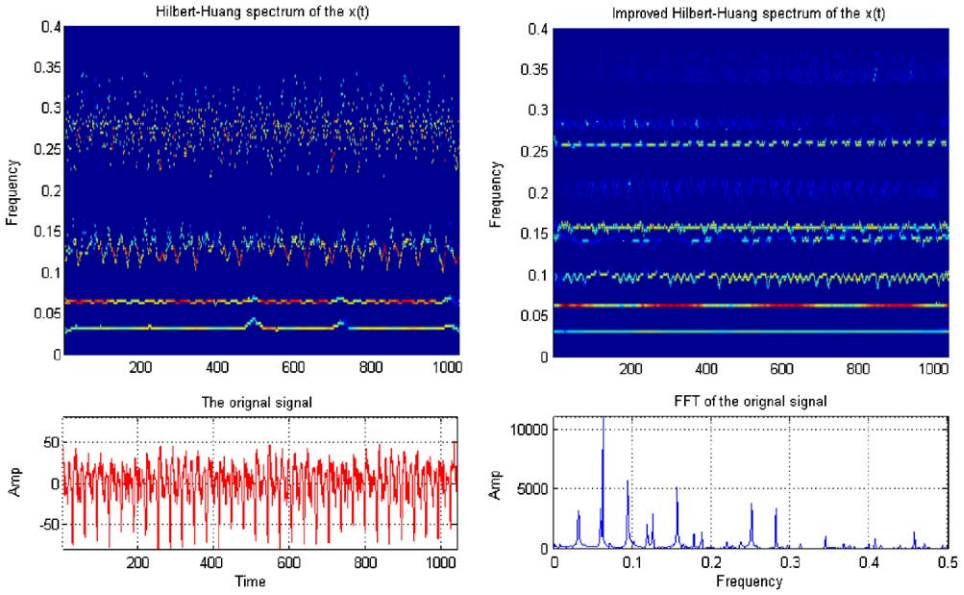

Fig. 13. The temporal waveform of a real vibration signal collected from serious rubbings, its FFT spectrum, and the results of the original HHT and the improved HHT.

results of the original HHT do not reveal these symptoms clearly; hence, it is not a qualified tool for rubbing type of signal analysis. Even worst, it may mislead the operators in making incorrect conclusion in rubbing fault diagnosis.

5. Conclusions

In this research, a novel time–frequency analysis method named as HHT has been introduced in detail. Three major shortcomings of the HHT have been put forward. They include the first obtained IMF may cover too wide a frequency range that may induce the HHT fails to produce an accurate frequency pattern for the inspected signal, the generation of undesirable pseudocomponents to the result that may cause misinterpretation to the result, and the loss of frequency components that have low energy content. To solve these shortcomings, an improved HHT has been suggested in this paper. The improved HHT utilizes the WPT to act as a preprocessor to decompose the raw signal into a set of narrow band signals. Then the EMD will be applied to these narrow band signals and extract sets of IMFs. After the IMFs have been obtained, then a simple prescreening process will be applied to retain those vital IMFs that are related to the raw signal. Both the original HHT and the improved HHT have been used to analyse simulated and experimental vibration signals generated from different degree of severity in rubbings. The results confirmed that the improved HHT is a promising tool for analysing rubbing vibration signal that exhibited nonlinear and non-stationary properties. Moreover, from the test results, two major symptoms have been identified for confirming the occurrence of rubbing fault and monitoring its deterioration. The two symptoms include (1) the high-frequency

components will change their frequency values periodically when rubbings occur, and (2) the low-frequency components will change their amplitudes periodically if the fault of rubbing becomes serious. Besides the fault of rubbing, other types of fault signals that exhibit nonlinear and non-stationary properties will be used to further verify the effectiveness of the improved HHT.

Acknowledgement

The work described in this paper was mainly supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 1100/02E), and partially supported by a China Postdoctoral Fund (Project No. 023212001).

References

[1] N.E. Huang, Z. Shen, S.R. Long, M. Wu, H. Shih, N. Zheng, C. Yen, C.C. Tung, H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis, *Proceedings of the Royal Society of London Series A—Mathematical Physical and Engineering Sciences* 454 (1998) 903–995.

[2] W.J. Wang, P.D. McFadden, Application of wavelets to gearbox vibration signals for fault detection, *Journal of Sound and Vibration* 192 (1996) 927–939.

[3] P. Tse, W.X. Yang, H.Y. Tam, Machine fault diagnosis through an effective exact wavelet analysis, *Journal of Sound and Vibration* 277 (2004) 1005–1024.

[4] N.E. Huang, Z. Shen, S.R. Long, A new view of non-linear water waves: the Hilbert spectrum, *Annual Review of Fluid Mechanics* 31 (1999) 417–457.

[5] M.E. Montesinos, J.L. Munoz-Cobo, C. Perez, Hilbert–Huang analysis of BWR neutron detector signals: application to DR calculation and to corrupted signal analysis, *Annals of Nuclear Energy* 30 (2003) 715–727.

[6] B. Yang, C.S. Suh, Interpretation of crack-induced rotor non-linear response using instantaneous frequency, *Mechanical Systems and Signal Processing* 18 (3) (2004) 491–513.

[7] J.N. Yang, Y. Lei, System identification of linear structures using Hilbert transform and empirical mode decomposition, *Proceedings of the 18th International Modal Analysis Conference, A Conference on Structural Dynamics*, San Antonio, TX, USA, Vol. 1, 2000, pp. 213–219

[8] R.L. Dequeiroz, K.R. Rao, Time-varying lapped transforms and wavelet packets, *IEEE Transactions on Signal Processing* 41 (1993) 3293–3305.

[9] Z. Peng, P. Tse, F. Chu, A comparison study of improved Hilbert–Huang transform and wavelet transform: application to fault diagnosis for rolling bearing, *Mechanical Systems and Signal Processing* 19 (2005) 974–988.

[10] X. Dai, Z. Jin, X. Zhang, Dynamic behavior of the full rotor/stop rubbing: numerical simulation and experimental verification, *Journal of Sound and Vibration* 251 (2002) 807–822.

[11] Z. Peng, F. Chu, P. Tse, Detection of the rubbing caused impacts for rotor-stator fault diagnosis using reassigned scalogram, *Mechanical Systems and Signal Processing* 19 (2005) 391–409.

[12] N.Q. Hu, M. Chen, X.S. Wen, The application of stochastic resonance theory for early detecting rubbing caused impacts fault of rotor system, *Mechanical Systems and Signal Processing* 17 (2003) 883–895.

[13] Z. Peng, Y. He, Q. Lu, F. Chu, Feature extraction of the rubbing caused impacts rotor system by means of wavelet analysis, *Journal of Sound and Vibration* 259 (2003) 1000–1010.

[14] F. Chu, Z. Zhang, Bifurcation and chaos in a rubbing caused impacts Jeffcott rotor system, *Journal of Sound and Vibration* 209 (1998) 1–18.