IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 22, 2017, accepted October 9, 2017, date of publication October 25, 2017, date of current version November 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2766232

Adaptive Mode Decomposition Methods and Their Applications in Signal Analysis for Machinery Fault Diagnosis: A Review With Examples

ZHIPENG FENG[1], (Senior Member, IEEE), DONG ZHANG[1], AND MING J. ZUO[2,3], (Senior Member, IEEE)

[1]School of Mechanical Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2]School of Mechatronics Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[3]Department of Mechanical Engineering, University of Alberta, Edmonton, AB T6G 2G8, Canada

Corresponding author: Ming J. Zuo (ming.zuo@ualberta.ca)

This work was supported in part by the National Natural Science Foundation of China under Grant 51475038 and in part by the Natural Sciences and Engineering Research Council of Canada under Grant RGPIN-2015-04897.

ABSTRACT Effective signal processing methods are essential for machinery fault diagnosis. Most conventional signal processing methods lack adaptability, thus being unable to well extract the embedded meaningful information. Adaptive mode decomposition methods have excellent adaptability and high flexibility in describing arbitrary complicated signals, and are free from the limitations imposed by conventional basis expansion, thus being able to adapt to the signal characteristics, extract rich characteristic information, and therefore reveal the underlying physical nature. This paper presents a systematic and up-to-date review on adaptive mode decomposition in two major topics, i.e., mono-component decomposition algorithms (such as empirical mode composition, local mean decomposition, intrinsic time-scale decomposition, local characteristic scale decomposition, Hilbert vibration decomposition, empirical wavelet transform, variational mode decomposition, nonlinear mode decomposition, and adaptive local iterative filtering) and instantaneous frequency estimation approaches (including Hilbert-transform-based analytic signal, direct quadrature, and normalized Hilbert transform based on empirical AM-FM decomposition, as well as generalized zero-crossing and energy separation) reported in more than 80 representative articles published since 1998. Their fundamental principles, advantages and disadvantages, and applications to signal analysis in machinery fault diagnosis, are examined. Examples are provided to illustrate their performance.

INDEX TERMS Adaptive mode decomposition, mono-component, instantaneous frequency, time-frequency representation, fault diagnosis.

I. INTRODUCTION

Signal analysis is a key step in machinery fault diagnosis. Effective extraction of signal features is helpful to well reveal the underlying physical nature of a phenomenon, thoroughly understand the dynamic characteristics of a system, and to evaluate its health condition, thereby providing convincing evidences for fault diagnosis. However, in both academic researches and engineering practices, real signals are usually highly intricate, because many sorts of modern machinery are often composed of mechanical, electrical and hydraulic systems. Their dynamic responses are a complex mixture of various dynamic phenomena, including mechanical vibrations, electrical oscillations, hydraulic fluctuations, their coupling effects, and the dynamic responses to external environmental excitations. As a consequence, their dynamic responses feature multi-components, high degree of complexity and various types of morphology. Such complicated dynamic responses contain rich information about the running condition and health status of machinery. Therefore, complicated signal analysis is a key research topic and plays an important role in machinery fault diagnosis.

To date, various signal analysis methods have been proposed for machinery fault diagnosis [1]–[5]. These include Fourier, wavelet (packet), lifting wavelet and multi-wavelet

transforms, as well as various time–frequency analysis methods such as Cohen and affine class distributions, atomic decomposition, time varying higher order spectra, and Hilbert-Huang transform [1]–[5]. Most of them are based on basis expansion. For example, Fourier series expansion lays a foundation for most spectral analysis methods, and wavelet basis expansion is the core of wavelet analysis. Traditional basis expansion methods have good merits, such as simplicity, uniqueness and symmetry, but they suffer from inflexibility and lack degree of freedom, because a priori knowledge on signals is needed to construct explicit bases and these bases are subject to orthogonality constraint. As such, traditional basis expansion methods are not adaptive enough for arbitrary complicated signal analysis.

Adaptive mode decomposition methods provide an effective approach to arbitrary complicated signal analysis. Such decomposition methods are truly data-driven and posterior. They neither need to construct any a priori basis to match the signal characteristic structure nor impose any constraints on the way to represent signals in time, frequency and joint time–frequency domains. They adapt to the transient feature, and highlight the local characteristics of signals. Therefore, they can adaptively extract the constituent oscillation modes of mono-component nature (without an explicit expression in closed form) which reflect the underlying oscillation property, from an arbitrary signal, and represent the signal as a superposition of several mono-components (and a residue). This mono-component decomposition nature enables accurate estimation of both the instantaneous frequency and the instantaneous amplitude of each constituent component. These instantaneous parameters provide insight into the frequency composition and time variability of signals. Based on the instantaneous frequency and instantaneous amplitude, a quality time–frequency representation can be constructed as a linear superposition of constituent components' time–frequency representations, which has high time–frequency resolution and is free from both inner and outer interferences, thus being effective in resolving the frequency contents and their time–frequency structure of arbitrary nonstationary complicated signals [6]–[26]. Thanks to the above merits, adaptive mode decomposition methods are highly adaptive to the complicated and various morphological contents, thus being effective in separating harmonic, impulsive and modulated components, and in extracting the dynamic features of a system as well, through analysis of the amplitude and frequency of each resultant mono-component and the time–frequency representation of the signal.

Machinery fault diagnosis usually relies on detecting the presence of certain fault frequencies and/or monitoring their time variability in terms of both frequency and amplitude. Therefore, it is necessary to accurately calculate the instantaneous frequency of target components. Adaptive mode decomposition can decompose an arbitrary complicated signal into its constituent components, thus meeting the mono-component requirement by instantaneous frequency calculation. The derived time–frequency representation can effectively identify the frequency contents of a signal and exhibit their time variability, for example, the rotating frequency and its harmonics in rotating machinery vibration signals. In particular, for gearboxes and rolling bearings, fault information is mainly carried by mono-components with instantaneous frequency fluctuates around the signal carrier frequency (gear meshing frequency and its harmonics in gearbox case, and resonance frequency in rolling bearing case). In-depth analysis of such sensitive mono-components (Fourier spectra of their instantaneous amplitude and instantaneous frequency) may reveal fault features in details, such as amplitude and frequency modulations characteristic of gear and rolling bearing faults.

In summary, adaptive mode decomposition methods can analyze arbitrary complicated multi-component signal more flexibly, and give insights into the signal nature from various perspectives, thus better extracting the rich information. They break through the limitations inherent with basis expansion based methods, and overcome the shortcomings of conventional methods, thus providing an effective approach to complicated signal analysis in machinery fault diagnosis.

Since Huang *et al.* [6] proposed the Hilbert-Huang transform (HHT) in 1998, adaptive mode decomposition has attracted more and more researchers' attention, and some fundamental outcomes have been obtained. Although the empirical mode decomposition (EMD) is effective in mono-component decomposition, it has some shortcomings, such as lack of mathematic formulation, susceptibility to mode mixing under singularities, instability under noise interferences, over/under fitting due to cubic spline interpolation. Inspired by the idea of adaptive mode decomposition, some new methods have been proposed to address the issues existing with EMD, such as: the local mean decomposition (LMD) by Smith [13] in 2006, the intrinsic time scale decomposition (ITD) by Frei and Osorio [14] in 2007, the local characteristic scale decomposition (LCD) by Zheng *et al.* [15] in 2013, for solving the mode mixing and over/under fitting problems, which follow the same mono-component sifting framework of EMD; as well as the Hilbert vibration decomposition (HVD) by Feldman [16]–[21] in 2006, the empirical wavelet transform (EWT) by Gilles [22] in 2013, the variational mode decomposition (VMD) by Dragomiretskiy and Zosso [23] in 2014, the nonlinear mode decomposition (NMD) by Iatsenko *et al.* [24] in 2015, and the adaptive local iterative filtering (ALIF) by Cicone *et al.* [25] in 2016, for a rigorous mathematic formulation and better robustness to noise, which separate mono-components by exploiting their amplitude modulation and frequency modulation (AM-FM) property or the filtration nature of EMD.

Instantaneous frequency is a key parameter to describe the physical nature of each mono-component obtained from aforementioned adaptive mode decomposition algorithms. The most widely used instantaneous frequency estimation approach is based on analytic signal via Hilbert transform (HT). However, it suffers from some drawbacks such as negative frequency values and frequency

fluctuation [27], [28]. To overcome these drawbacks, some approaches have also been proposed, such as energy separation (ES) [29], [30], generalized zero-crossings (GZC), empirical AM-FM decomposition, direct quadrature (DQ), and normalized Hilbert transform (NHT) [28].

To date, many articles on adaptive mode decomposition have been published. Investigations on these methods and their applications are still ongoing. Many new research results, in terms of both adaptive mode decomposition algorithms, instantaneous frequency estimation approaches, and their applications, are being reported from various fields every year. They have also been applied to analysis of complicated signals in machinery fault diagnosis [31]–[93]. A systematic review on adaptive mode decomposition methodology and its application in machinery fault diagnosis would benefit researchers in this field. However, reported review papers [1]–[4] do not fully cover this topic. Lei *et al.* [5] made a review towards this direction, but they focused on EMD and its ensemble version only. There lacks an extensive review on all adaptive mode decomposition methods (including EMD, LMD, ITD, LCD, HVD, EWT, VMD, NMD, and ALIF) and instantaneous frequency estimation approaches (such as HT, ES, GZC, DQ, and NHT). Moreover, regarding the characteristics of intricate signals encountered in machinery fault diagnosis, how to exploit the merits of these methods, and effectively extract the meaningful features, still deserves further investigation in-depth.

This paper aims to provide a comprehensive study and a systematic review of adaptive mode decomposition methods, thus offering guidance for researchers who are interested in this methodology. Extensive state-of-the-art adaptive mode decomposition methods are summarized, including their ideas, algorithms, and applications in machinery fault diagnosis. An up-to-date review of the existing literature and some insights into studies of the latest adaptive mode decomposition methods are presented. For readers to quickly understand the underlying idea and/or mathematic rationale of adaptive mode decomposition, we divide these methods into two major classes: one is adaptive mode decomposition algorithms, and the other is instantaneous frequency estimation approaches. For each method, we introduce its mathematical principles, illustrate its merits via analysis of a representative synthetic signal, summarize its pros and cons, review its applications in machinery fault diagnosis, and point out future research directions. Such a review would guide engineers to select properly an adaptive mode decomposition method according to its suitability, signal characteristics and analysis demand, and motivate or inspire researchers to improve the existing signal feature extraction methods and explore new ones, thus addressing the complicated signal analysis issue in machinery fault diagnosis. The existence of large body of literature developed over the past two decades makes it unrealistic to review each and every article published in this field. Therefore, we will focus on recent key advances in adaptive mode decomposition methods and their typical applications in machinery fault diagnosis.

Hereafter, this paper is organized as follows. Firstly, the fundamentals of adaptive mode decomposition are introduced in Section II. Then, adaptive mode decomposition algorithms and instantaneous frequency estimation approaches are reviewed in Sections III and IV respectively. In Section V, some application examples are presented to illustrate the potential of typical adaptive mode decomposition methods in machinery fault diagnosis. Finally, in Section VI, the pros and cons of these adaptive mode decomposition methods are summarized, and some application prospects and remaining research issues in machinery fault diagnosis are pointed out.

II. GENERAL PRINCIPLE

Complicated signals are usually composed of multi-components, and in many cases, they are nonstationary and nonlinear, i.e. each constituent component exhibit time variability, in terms of amplitude, phase and/or frequency. Each component can be considered as an amplitude modulation and frequency modulation (AM-FM) oscillatory mode. Hence, an arbitrary complicated multi-component signal can be modeled as a superposition of several AM-FM components

$$\begin{aligned} x(t) &= \sum_{i=1}^{N} c_i(t) = \sum_{i=1}^{N} a_i(t)\cos[\phi_i(t)] \\ &= \sum_{i=1}^{N} a_i(t)\cos\left[\omega_c t + \int \omega_i(t)dt\right], \end{aligned} \tag{1}$$

where $a_i(t)$ is the instantaneous amplitude, $\phi_i(t)$ the instantaneous phase, ω_c the carrier frequency, and $\omega_c + \omega_i(t) = \dot{\phi}_i(t)$ the instantaneous frequency. In this paper, the instantaneous amplitude $a_i(t)$ and instantaneous frequency $\omega_c + \omega_i(t)$ are assumed to be slowly varying compared to the carrier frequency ω_c.

To study the properties of such nonstationary signals, it is necessary to access the instantaneous parameter of each constituent component, including the instantaneous amplitude and instantaneous frequency. However, the instantaneous frequency is meaningful for single frequency component only, and thus is defined based on mono-component. Hence, an arbitrary multi-component signal is not automatically ready to calculate the instantaneous frequency, and it has to be decomposed into mono-components, i.e. individual oscillatory modes are separated from each other, each with a physically meaningful instantaneous frequency. In order to enable the instantaneous frequency estimation, conditions are defined to guarantee the separated individual component be a mono-component. According to the conditions, adaptive mode decomposition algorithm can be designed, and an arbitrary complicated signal can then be decomposed into several mono-components (and a residue)

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t), \tag{2}$$

where $c_i(t)$ is the mono-component, and $r_n(t)$ is the residue and can be either a mean trend or a constant.

Once mono-components are obtained, their instantaneous amplitude, instantaneous phase and instantaneous frequency can be estimated, see Section IV. Given these instantaneous parameters, the signal time variability can be studied in-depth. For example, transient events can be detected via studying the local changes in these parameters, and the AM and/or FM properties can be revealed by analyzing the instantaneous amplitude and instantaneous frequency respectively.

Meanwhile, the frequency contents of a given nonstationary signal and the time variability of each constituent component can also be extracted via time–frequency analysis. Given the instantaneous amplitude $a_i(t)$ and instantaneous frequency $\omega_c + \omega_i(t)$ of each mono-component $c_i(t)$, the time–frequency representation can be derived as

$$\mathrm{TFR}(t, \omega) = \sum_{i=1}^{n} a_i(t)\delta\{\omega - [\omega_c + \omega_i(t)]\}, \tag{3}$$

where $\delta(\cdot)$ is Dirac delta function. Such time–frequency representation is free from both outer and inner interferences. More importantly, it has fine time–frequency resolution and good readability, because: it is a linear superposition of the time–frequency representation of constituent mono-components rather than a double integral involving quadratic terms of multiple components, and the instantaneous frequency is defined as the local derivative of instantaneous phase which emphasizes the local properties of signals.

Since mono-component decomposition is a key to success in instantaneous frequency estimation, and accurate instantaneous frequency estimation is essentially important to reveal the frequency contents and their time variability of nonstationary complicated signals, we review the adaptive mode decomposition algorithms and instantaneous frequency estimation approaches in the following Sections III and IV respectively.

III. ADAPTIVE MODE DECOMPOSITION ALGORITHMS

For effective analysis of complicated signals, inspired by the idea of EMD, various adaptive mode decomposition algorithms have been proposed to decompose intricate multi-component signals into constituent mono-components. Typical algorithms include EMD, LMD, ITD, LCD, HVD, EWT, VMD, NMD and ALIF, to be reviewed in this section.

A. EMPIRICAL MODE DECOMPOSITION AND ENSEMBLE EMPIRICAL MODE DECOMPOSITION

1) PRINCIPLE

The EMD proposed by Huang *et al.* [6], [7] can adaptively decompose a complicated multi-component signal into constituent mono-components. This algorithm recursively detects local minima and maxima in a signal, data fits the lower and upper envelopes by interpolation of these extrema in local characteristic time scale, removes the instantaneous mean of the lower and upper envelopes as a "low-pass" centerline, thus separating the high-frequency components as intrinsic mode functions (IMFs, mono-component in nature), and continues recursively on the remaining "low-pass" centerline. Via such iterative sifting, any signal can be decomposed into a series of IMFs which satisfy the mono-component requirements by instantaneous frequency calculation: (1) in the whole time span, the number of extrema and the number of zero crossings must either equal or differ at most by one; and (2) at any time, the instantaneous mean of the upper and lower envelopes is zero. The first condition is similar to the traditional narrow band requirements for a stationary Gaussian process. The second condition modifies the classical global requirement to a local one, and makes the instantaneous frequency avoid the undesired fluctuations induced by asymmetric waveforms.

For a real signal $x(t)$, the EMD procedure is detailed as follows.

Step 1: Find the local minima and the local maxima of $x(t)$.

Step 2: Construct the lower envelope $L(t)$ and the upper envelope $U(t)$ of $x(t)$ respectively, by cubic spline interpolation to the local minima and the local maxima.

Step 3: Calculate the instantaneous mean of the lower and upper envelopes $m(t) = [L(t) + U(t)]/2$.

Step 4: Construct a prototype IMF $h(t) = x(t) - m(t)$.

Step 5: If $h(t)$ satisfies the stop criteria for IMF sifting, then set it as an IMF $c(t) = h(t)$. Otherwise, repeat steps 1-5 on $h(t)$.

Step 6: Construct a residual signal $r(t) = x(t) - c(t)$.

Step 7: If $r(t)$ satisfies the stop criteria for EMD, then set $r(t)$ as the final residual signal, and terminate the EMD process. Otherwise, repeat steps 1-7 on $r(t)$.

The IMF sifting procedure, steps 1-5, eliminates riding waves and makes the profile of prototype IMFs more symmetric with respect to zero. However, over-sifting may lead to loss of amplitude variation and physical meaning. In order to guarantee proper sifting and meaningful IMF, several stop criteria for IMF sifting have been proposed. To name a few for example, a Cauchy-type criteria is set according to the standard deviation of two consecutive prototype IMFs

$$\sigma = \frac{\|h_{i+1}(t) - h_i(t)\|_2}{\|h_i(t)\|_2}. \tag{4}$$

When the standard deviation reaches a predefined threshold usually set between 0.2 and 0.3, stop the inner sifting loop steps 1-5 [6].

An S-number criterion is based on the numbers of extreme and zero-crossings. For a prototype IMF, when the number of extrema equals the number of zero-crossings for predefined (usually between 3 and 5) steps of successive sifting, the inner sifting loop can be stopped [7], [8].

A combined global–local criterion guarantees globally small fluctuations in the instantaneous mean while considering locally large excursions. It evaluates the amplitude of instantaneous mean $m(t)$ in comparison with the amplitude

of corresponding prototype IMF $a(t) = U(t) - L(t)$

$$\sigma(t) = \frac{m(t)}{a(t)}. \tag{5}$$

If $\sigma(t) < \theta_1$ for some prescribed fraction $(1 - \alpha)$ of the total duration, and $\sigma(t) < \theta_2$ for the remaining fraction, stop the sifting procedure. A default setting is $\alpha = 0.05$, $\theta_1 = 0.05$ and $\theta_2 = 10\theta_1$ [9].

A local stop criterion fixes the sifting number, thus avoiding pseudo local extrema due to over-sifting in case of local wiggles in prototype IMFs. The optimal sifting number is often set to 10 [10], [11].

For the whole EMD procedure, when the residual $r(t)$ becomes a trend, i.e. it has one local extremum at most, stop the outer iteration loop steps 1-7.

The EMD essentially projects a signal onto the time-frequency plane, makes each projection a mono-component, and preserves the time-varying characteristics of each component, thus being able to calculate the instantaneous frequency. It separates IMFs in a frequency order from high to low, and the result has a dyadic frequency band decomposition property when applied to white noise. The decomposition by EMD is complete, adaptive, and almost orthogonal in applications [6].

Although the EMD is well known and widely used, it suffers from a major drawback of possible mode mixing (a phenomenon that disparate scales appear in one IMF, or a coherent component of a similar scale resides in more than one IMFs), which is often caused by intermittences in signals. In order to overcome this drawback, a noise assisted EMD, the ensemble empirical mode decomposition (EEMD), was proposed [11].

The EEMD is inspired by the dyadic filter bank nature of the EMD when applied to white noise [10], [11]. White noise of finite amplitude is added to the signal to provide a uniform reference frame in the time–frequency space, perturb the signal in the neighborhood of its true solutions, thereby force the ensemble to exhaust all possible solutions in the EMD sifting process, and enable the signal components of different scales to collate in proper IMFs. In the ensemble mean of sufficient trials, the noise will be averaged out since it is different in separate trials. The EEMD involves the following steps:

Step 1: Add a white noise series to the signal.

Step 2: Decompose the signal with added white noise into IMFs using the traditional EMD.

Step 3: Return to step1 and redo steps 1-2 for a predefined number of iterations, but with different white noise series each time.

Step 4: Obtain the ensemble means of corresponding IMFs as the final result.

In the EEMD, the number of trials N in the ensemble and the amplitude of the added noise a are two key parameters to be carefully selected. When the number of ensembles approaches infinity, the EEMD produces true decomposition. In practice, the number of ensembles is limited. Therefore, the resultant IMFs are inevitably contaminated by added noise. The standard deviation of error is given below

$$e = \frac{a}{\sqrt{N}}. \tag{6}$$

To reduce the error, small noise amplitude is preferred. However, if the noise amplitude is too small, it may not introduce sufficient change of extrema that EMD relies on. Hence the noise amplitude should not be too small. Under this condition, the noise effect can be reduced to a negligible level with the increased number of trials. To make the EEMD effective, the amplitude of noise is suggested to be 0.2 times the standard deviation of the signal, and the number of trials in an order of a few hundreds [11].

To address the issue of residual noise in resultant IMFs, a complementary EEMD was developed. For a white noise, both its positive and negative versions are added to data as complementary trials. The residue of added noise can be cancelled out by averaging complementary ensemble IMFs [12].

2) ILLUSTRATION

To illustrate the performance of adaptive mode decomposition algorithms, we generate a synthetic signal according to the following equation

$$\begin{aligned} x(t) = {} & \cos[2\pi f_{\text{carrier1}} t + 125\cos(2\pi f_{\text{FM}} t)] \\ & + [2 + \cos(2\pi f_{\text{AM}} t)]\cos(2\pi f_{\text{carrier2}} t) + n(t). \end{aligned} \tag{7}$$

It consists of two true components and a white Gaussian noise $n(t)$ at a signal-to-noise ratio of 20 dB. One true component is a sinusoidal frequency modulation (FM) component at a modulating frequency of $f_{\text{FM}} = 0.5$ Hz and riding on a carrier frequency of $f_{\text{carrier1}} = 137.5$ Hz, and the other is a sinusoidal amplitude modulation (AM) component at a modulating frequency of $f_{\text{AM}} = 0.5$ Hz and riding on a carrier frequency of $f_{\text{carrier2}} = 30$ Hz. The instantaneous frequency of the sinusoidal FM component changes over time nonlinearly. This signal is simple, but is representative because of its multi-component nature and nonstationarity. More importantly, it simulates the common yet typical phenomena often encountered in machinery fault diagnosis, for example, the modulation features characteristic of rolling bearing and gear faults, and the time variability of major frequency components in rotating machinery vibration signals during variable speed conditions. It will be used to illustrate each adaptive mode decomposition algorithm in the following sections.

Fig. 1 (a) shows the IMFs obtained from the EMD. We use the combined global–local stop criterion for IMF sifting, and set $\alpha = 0.05$, $\theta_1 = 0.05$ and $\theta_2 = 10\theta_1$, according to the recommendation in [9]. The two constituent components are well separated and clearly identified. The first two correspond to the sinusoidal FM and AM components respectively. The rest are not the true constituent components of the signals. They are possibly caused by the approximation error of cubic spline fitting and the end effect, but their amplitudes are small and thus are negligible. The instantaneous frequency

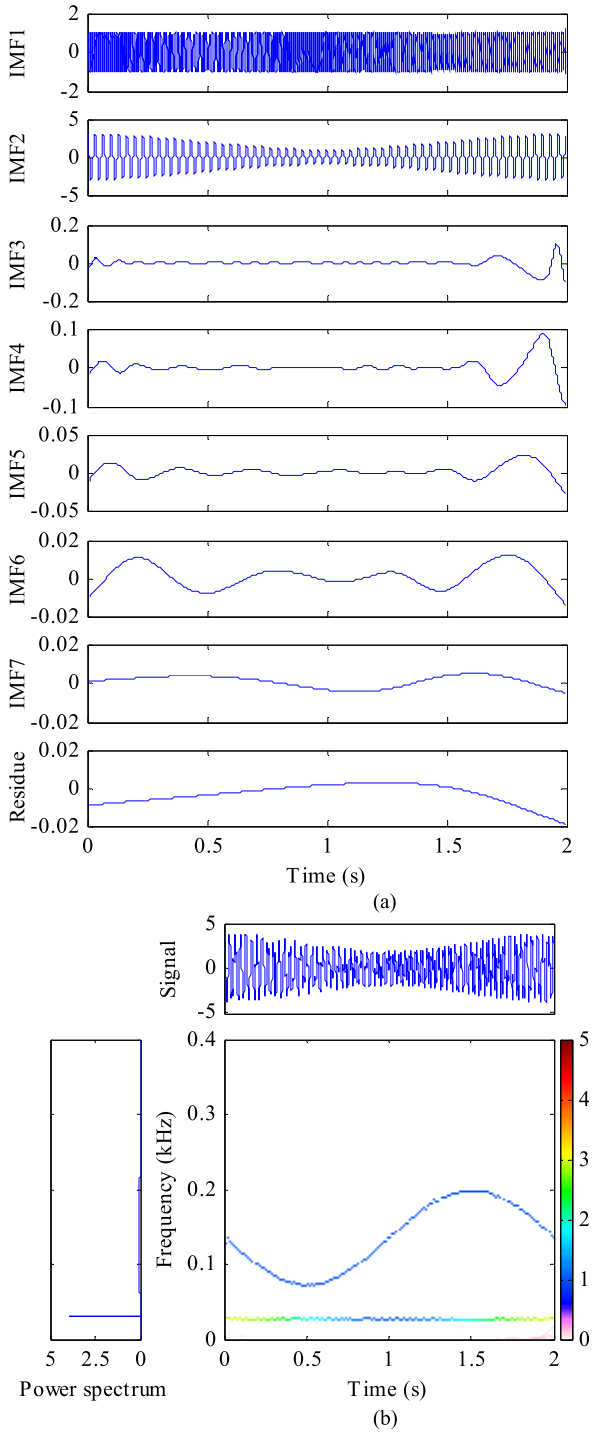

FIGURE 1. EMD analysis result. (a) IMFs and residue; (b) Hilbert energy spectrum.

of each IMF is calculated using the Hilbert transform based analytic signal approach (the same in Part Illustration for other adaptive mode decomposition algorithms if not stated specifically). Fig. 1 (b) shows the Hilbert energy spectrum of synthetic signal. The true components are resolved in fine details. The sinusoidal FM component with highly nonlinear instantaneous frequency but constant amplitude is clearly exhibited as a sinusoidal curve on the time–frequency plane. The time-variant amplitude of sinusoidal AM component is also revealed by time–frequency energy distribution magnitude (represented by color) of the associated constant carrier frequency. Their instantaneous frequency trajectories exactly follow the theoretical ones, and their amplitudes also match with the true ones.

Fig. 2 shows the EEMD analysis result. In the EEMD, we set the amplitude of noise to be 0.2 times the signal standard deviation, and the number of trials to be 100, according to the suggestions in [11]. The two constituent components are well separated. In Fig. 2 (a), IMF 1 and 2 link to the sinusoidal FM and AM components respectively. Some irrelevant IMFs 3-8 are also generated due to EMD error and added assisting noise, but their amplitude is small and can be neglected. Accordingly, in the Hilbert energy spectrum, Fig. 2 (b), the two frequency components and their time variability are clearly exhibited.

3) APPLICATION REVIEW

Many research papers on the application of the EMD and EEMD to machinery fault diagnosis have been published since 2000. To name a few for example, Cheng *et al.* [31]–[33] applied the EMD to fault diagnosis of rolling bearings, gears and rotors. They proposed a rolling bearing fault diagnosis method based on marginal Hilbert spectrum of the envelope signal obtained from wavelet decomposition. They found that impulses caused by gear fault could be detected in the instantaneous energy, and showed that the EMD was useful to separate the modulation components due to rotor-stator rub-impact. To avoid mode mixing and reduce computational cost, Zheng *et al.* [34] developed a partly EEMD based on complementary EEMD and by merits of permutation entropy in detecting random noises and intermittences. The signal is decomposed via complementary EEMD until the permutation entropy of residue is higher than a predefined threshold. Then the residue is further decomposed through EMD. They detected rotor-stator rubbing using the partly EEMD. Peng *et al.* [35], [36] proposed an improved Hilbert-Huang transform and extracted fault symptoms of rolling bearings and rotors in time–frequency domain. They used the wavelet packet transform as a preprocessor to decompose a signal into a set of narrow band signals, thus avoiding the deficiency of wide ranging frequency band of the first IMFs. The IMFs were selected via correlation analysis of the IMF with the raw signal, thus eliminating pseudo IMFs due to end effects. Loutridis [37] used the EMD to examine gear vibration signals, and found that the IMF energy and instantaneous frequency can be used to detect the gear root crack. Liu *et al.* [38] revised the EMD by B-spline fitting, and applied it to gearbox fault diagnosis. Ricci and Pennacchi [39] proposed an index for automatic IMF selection, which is a linear combination of the periodicity degree and absolute skewness of the IMF. They illustrated its effectiveness by applying it to gearbox fault diagnosis. Georgoulas *et al.* [40] extracted features from

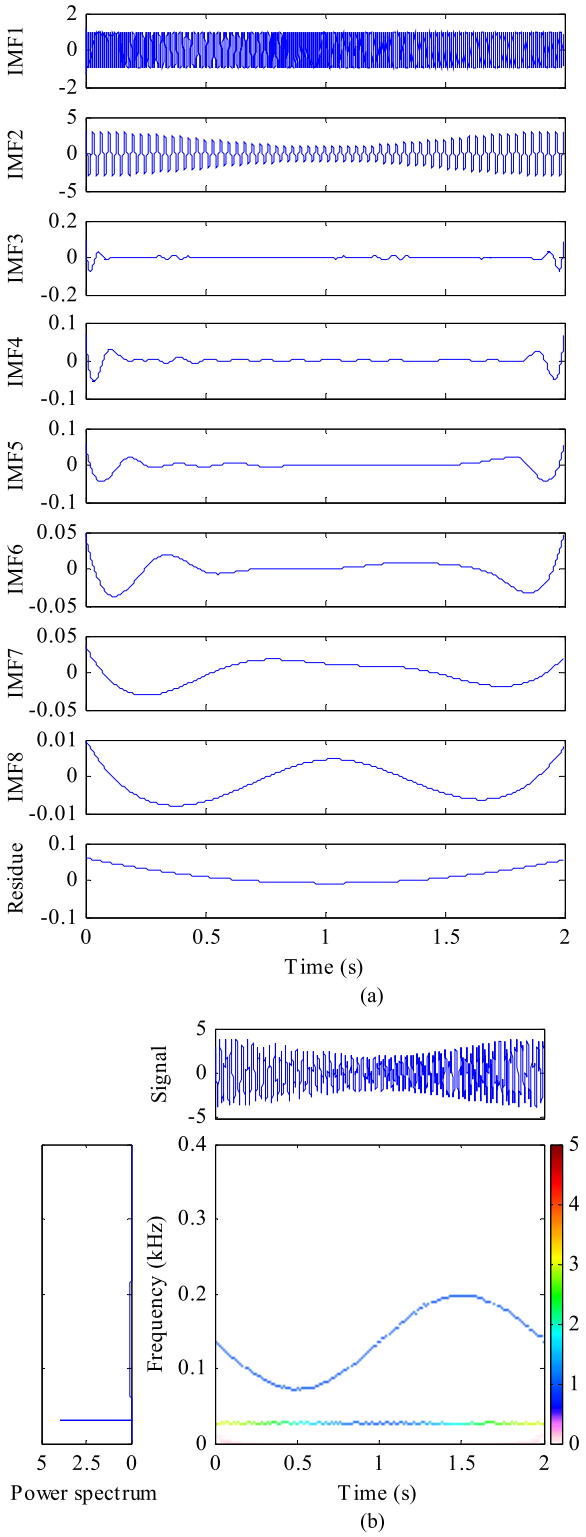

FIGURE 2. EEMD analysis result. (a) IMFs and residue; (b) Hilbert energy spectrum.

EMD based IMFs with highest kurtosis values, including the instantaneous frequency level and the spread of IMF, for rolling bearing anormaly detection. Feng *et al.* [41] exploited the mono-component decomposition capability of EMD, and proposed a demodulation analysis method based on sensitive IMFs, thus addressing the spectral complexity issue due to multiple modulation sources in planet bearing fault diagnosis. Lei *et al.* [42]–[44] proposed a correlation-based criterion for sensitive IMF selection to improve EEMD based Hilbert-Huang transform, and further presented an algorithm to adaptively select the sifting number for each IMF and to determine the magnitude of added noise according to the sensitivity of components to noise. Stronger noise and larger sifting number are adopted to extract higher frequency IMFs, while weaker noise and smaller sifting number are employed for lower frequency IMFs. They validated the algorithm using simulation data, and applied it to rotating machinery fault diagnosis. Zhang *et al.* [45] improved the computational efficiency of EEMD by adding band limited noise and applied it to bearing fault diagnosis. Žvokelj *et al.* [46] applied the EEMD and principal component analysis to bearing fault diagnosis. To avoid the intricate sidebands in Fourier spectra, Feng *et al.* [47] proposed a joint amplitude and frequency demodulation analysis idea for wind turbine planetary gearbox fault diagnosis, by exploiting the merits of EEMD in mono-component decomposition and the advantages of energy separation algorithm in instantaneous amplitude and instantaneous frequency estimation. Wang *et al.* [48] presented an enhanced EEMD and applied it to rolling bearing fault diagnosis. They fused successive IMFs with higher spectral coherence similar characteristics into a new IMF, thereby addressing the possible mode mixing issue with EEMD. Lei *et al.* [5] conducted a systematic review on EMD and EEMD, together with their applications in rotating machinery fault diagnosis, including rolling bearings, gears, and rotors. Readers can refer to [5] for a comprehensive summary of pertinent references.

4) REMARKS

EMD and EEMD do not need any priori basis to match the signal characteristic, but extracts adaptively the intrinsic fluctuation modes inherent in signals by means of numerical approximation. The derived time–frequency representation offers fine time–frequency resolution and is free from both inner and outer interferences. These advantages make it effective in resolving the time varying structure of signal components. However, the EMD and EEMD lack rigorous mathematical formulation. A higher sampling frequency is preferred. EMD and EEMD fit the upper and lower envelopes based on extrema, and hence they need a fair amount of oversampling to correctly identify extrema for fine interpolation. For signals with instantaneous frequency trajectory crossings, mode mixing is inevitable. The EMD and EEMD usually generate more than one IMF. Some researchers proposed methods to select a subset of relevant IMFs [35], [36], [39], [49], [50], whereas others combined IMFs together in order to ease the comparison of signal contents over repeated acquisitions [51]–[53]. How to effectively select IMFs sensitive to fault still lacks a generally effective criterion.

B. LOCAL MEAN DECOMPOSITION

1) PRINCIPLE

Smith [13] proposed the local mean decomposition (LMD) to estimate the instantaneous frequency and instantaneous amplitude of signals. It decomposes a signal into the product of amplitude modulation and frequency modulation (mono-component in essence), and then separates iteratively the frequency modulation from amplitude modulation. LMD obtains mono-components via data smoothing methods, rather than cubic spline fitting used in EMD. Unlike the IMFs obtained from EMD which do not contain oscillations without zero-crossings between successive extrema, the product functions (PFs) derived from LMD may well contain oscillations which do not cross zero. Therefore, it may retain more of the frequency and amplitude variations in signals than the EMD does. For a signal x(t), the procedure of LMD is detailed as follows.

Step 1: Find the local minima and the local maxima of $x(t)$, and denote them as $n(k_i)$, where $k_i = k_1, k_2, \cdots$ is the time index of extrema.

Step 2: Calculate the local mean of successive maxima and minima

$$m(t) = \frac{1}{2}[n(k_i) + n(k_{i+1})], \tag{8}$$

and the local magnitude of successive maxima and minima

$$a(t) = \frac{1}{2}|n(k_i) - n(k_{i+1})|, \tag{9}$$

where $t \in [k_i, k_{i+1})$.

Step 3: Interpolate the local mean and local magnitude values between successive extrema with straight lines.

Step 4: Construct a continuous local mean function $\hat{m}(t)$ and an amplitude function $\hat{a}(t)$ by smoothing the interpolated local mean and local magnitude via moving average weighted by the time-lapse between successive extrema.

Step 5: Construct a prototype PF

$$h(t) = x(t) - \hat{m}(t), \tag{10}$$

and an FM signal

$$s(t) = \frac{h(t)}{\hat{m}(t)}. \tag{11}$$

Step 6: If $\hat{m}(t)$ is close to 1, set $s(t)$ as a purely normalized FM. Otherwise, let $h(t) = s(t)$, and repeat steps (1) through (6) on $h(t)$.

Step 7: Calculate the instantaneous amplitude of PF

$$a(t) = \prod \hat{a}(t), \tag{12}$$

and construct the PF

$$c(t) = a(t)s(t). \tag{13}$$

Step 8: Construct the residual signal

$$r(t) = x(t) - c(t). \tag{14}$$

If $r(t)$ becomes monotonic, then LMD terminates. Otherwise, repeat steps 1-8 on $r(t)$.

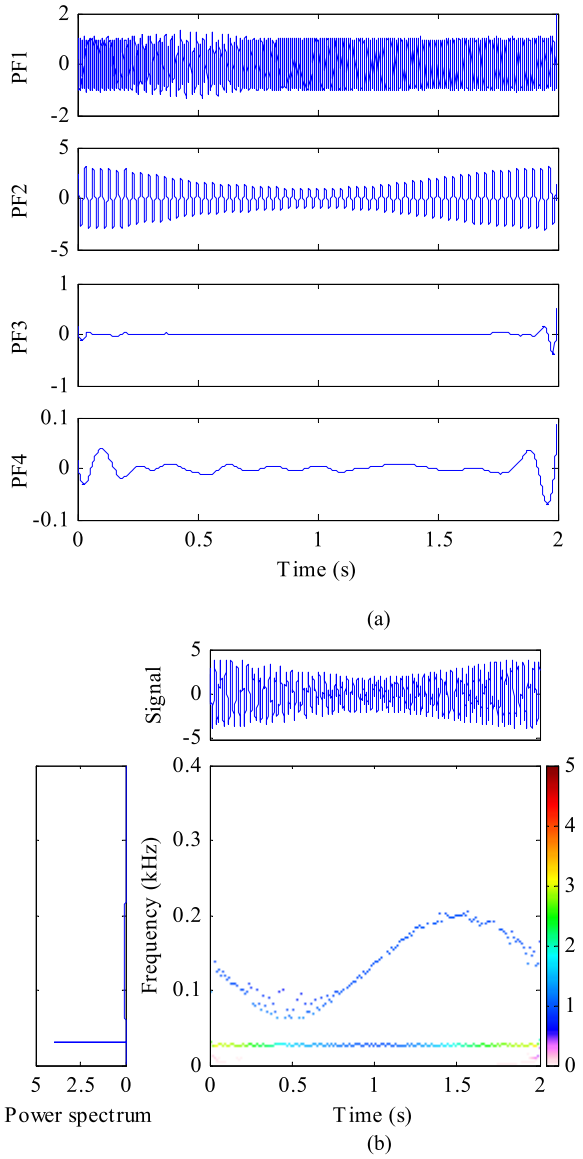

FIGURE 3. LMD analysis result. (a) PFs; (b) Time-frequency distribution.

The LMD share similar properties as the EMD, i.e. it separates the PFs in a frequency order from high to low, and exhibits a dyadic frequency band decomposition property when applied to white noise [54].

2) ILLUSTRATION

We apply the LMD to the synthetic signal in Section 3.1. Fig. 3 (a) shows the first four PFs obtained from the LMD. The first two correspond to the sinusoidal FM and AM components respectively. The other two likely result from the decomposition error. Their magnitudes are small, thus being negligible. Fig. 3 (b) shows the time–frequency distribution. It identifies the time–frequency structure of the sinusoidal AM component in a fine time–frequency resolution. For the sinusoidal FM component, when its instantaneous frequency is almost linear around 1 s, the time–frequency distribution concentrates around the true instantaneous frequency.

However, when its instantaneous frequency exhibits highly nonlinearity around 0.5 s and 1.5 s, the time–frequency distribution deviates from the true one. In particular, when the instantaneous frequencies of the two components are close, around 0.5 s, distortion occurs to the waveform of PF1. These flaws may mislead further analysis.

3) APPLICATION REVIEW

Although the LMD is a relatively new method, it has been applied to fault diagnosis of machinery. For example, Yang *et al.* [54] found the wavelet like frequency decomposition property of LMD, thereby proposed an ensemble improvement version to address the mode mixing issue of the LMD, and applied it to rotor fault diagnosis. Cheng *et al.* [55] used the LMD to diagnose bearing and gear faults. Wang *et al.* [56], [57] used the extrema on the ends to predict the local mean function and local amplitude on the ends, thus alleviating end effects. They accordingly proposed a new strategy to select the step size in moving average for estimating the local mean function and the amplitude function. They applied the improved local mean decomposition method to fault diagnosis of rotors and gearboxes. Liu *et al.* [58] applied LMD to analyze wind turbine gearbox vibration signals, and selected the impulsive PFs for further analysis and fixed-shaft gear fault diagnosis. Considering the AM-FM feature of planetary gearbox vibration signals, Feng *et al.* [59] utilized LMD to decompose signals into PFs, and selected sensitive PFs with an instantaneous frequency around the gear meshing frequency or harmonics, for quality joint amplitude and frequency demodulation analysis and further gear fault feature extraction. Zheng *et al.* [60] selected the meaningful PF obtained from LMD based on mutual information entropy, and used the generalized morphological fractal dimensions of selected PF to detect gear faults. Liu *et al.* [61] used second generation wavelet transform for signal denoising, and then decomposed the denoised signal into PFs to select a sensitive component. Through spectrum analysis of the selected PF, they detected gearbox and locomotive rolling bearing faults. These works imply the potential of LMD in vibration feature extraction for machinery fault diagnosis.

4) REMARKS

The LMD suffers the same shortcomings of possible mode mixing as the EMD does in resolving the time–frequency structure of signals with close instantaneous frequency trajectories and instantaneous frequency trajectory crossings. Moreover, the smoothing and step size have a significant effect on the decomposition, thus it is necessary to properly select these two parameters according to signal characteristics. Nevertheless, LMD provides a new way to decompose multi-component signals into mono-components. With the LMD, the mono-components are obtained via data smoothing methods, rather than through cubic spline fitting used in the EMD. Therefore the LMD may retain more of the frequency and amplitude variations in signals than the EMD does. The instantaneous frequency estimation approach based on the normalized frequency modulated signal can effectively avoid the possible distortion error caused by the amplitude modulation effect, and ensure a positive instantaneous frequency.

C. INTRINSIC TIME-SCALE DECOMPOSITION

1) PRINCIPLE

Frei and Osorio [14] proposed a mono-component decomposition method for complicated signal analysis, called intrinsic time–scale decomposition (ITD). The extrema in signal waveforms imply the existence of oscillations. The constituent mono-component is termed proper rotation component (PRC) in ITD, which has strictly positive values at all local maxima and strictly negative values at all local minima, and is suitable to calculate the instantaneous frequency and instantaneous amplitude. A PRC is actually a riding wave with highest frequency on a baseline. In order to meet the necessary requirement on PRCs, i.e. preserving the monotonicity of the residual signal between adjacent extrema, the baseline is constructed via linear transform based on extrema, so that the characteristics of the raw signal can be transferred to the baseline and the residual signal, and the temporal information of critical points is precisely preserved, with a temporal resolution equal to the time scale of the occurrence of extrema. In each decomposition, given the baseline, a PRC can be obtained directly and immediately by subtracting the baseline from the input signal.

For a signal $x(t)$, define a baseline extraction operator L to separate the lower frequency baseline signal, i.e. $Lx(t)$ represents the instantaneous mean of the signal, written as $L(t)$. Define $Hx(t) = x(t) - L(t)$ the proper rotation component, written as $H(t)$. Then the signal can be decomposed as $x(t) = L(t) + H(t)$. Based on above definitions, the ITD algorithm is detailed as follows.

Step 1: Find the extrema of the signal $x(t)$, written as x_k, and the corresponding occurrence time instant τ_k, where $k = 0, 1, 2, \cdots$. Without loss of generality, let $\tau_0 = 0$.

Step 2: Suppose the operators $L(t)$ and $H(t)$ are given over the interval $[0, \tau_k]$, and the signal $x(t)$ exists on the interval $[0, \tau_{k+2}]$, then on the interval $(\tau_k, \tau_{k+1}]$ between adjacent extrema x_k and x_{k+1}, the piecewise baseline extraction operator is defined as

$$Lx(t) = L(t) = L_k + \frac{L_{k+1} - L_k}{L_{k+2} - L_k}[x(t) - x_k], \quad t \in (\tau_k, \tau_{k+1}], \tag{15}$$

where

$$L_{k+1} = \alpha[x_k + \frac{\tau_{k+1} - \tau_k}{\tau_{k+2} - \tau_k}(x_{k+2} - x_k)] + (1 - \alpha)x_{k+1}, \tag{16}$$

and $0 < \alpha < 1$, usually $\alpha = 0.5$.

Step 3: The operator for extracting PRC is defined as

$$H(t) = Hx(t) = x(t) - Lx(t) = x(t) - L(t). \tag{17}$$

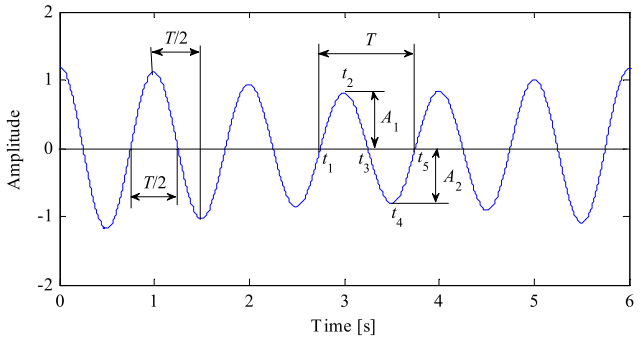

FIGURE 4. Definition of single wave, half wave and amplitude.

Take the baseline $L(t)$ as the input signal $x(t)$, and repeat steps 1-3, until the baseline becomes a monotonic function or a constant. Eventually, the raw signal can be decomposed into PRCs and a trend

$$x(t) = \sum_{i=1}^{p} H_i(t) + L_p(t), \tag{18}$$

where p is the number of the obtained PRCs.

Given a mono-component PRC, its instantaneous amplitude and frequency/phase can be extracted using a single wave based method piecewise. Take Fig. 4 as an example to introduce the definitions on single wave, half wave, and amplitude. Zero up-crossing refers to the zero crossings when the signal magnitude is increasing, see the points corresponding to time instants t_1 and t_5. Zero down-crossing stands for the zero crossings when the signal magnitude is decreasing, for example the point at time t_3. A single wave means the waveform between two adjacent zero up/down-crossings, for instance the one between time instants t_1 and t_5. A half wave denotes the signal waveform between any two adjacent zero crossings, for example the waveform between time instants t_1 and t_3. t_2 corresponds to the maximum of the positive half wave A_1. t_4 is the time when the minimum of the negative wave $-A_2$ occurs. A monotonic interval refers to the duration between any two adjacent extrema, for example $[t_2, t_4]$.

Instantaneous amplitude is defined based on the half wave. It is the signal amplitude at the extrema between two adjacent zero crossings, and is constant for a half wave

$$A_1(t) = A_2(t) = \begin{cases} A_1, & t \in [t_1, t_3) \\ -A_2, & t \in [t_3, t_5). \end{cases} \tag{19}$$

Instantaneous phase is defined based on the single wave, so as to guarantee the monotonicity of PRCs

$$\varphi(t) = \begin{cases} \arcsin\left[\dfrac{x(t)}{A_1}\right], & t \in [t_1, t_2) \\ \pi\text{-}\arcsin\left[\dfrac{x(t)}{A_1}\right], & t \in [t_2, t_3) \\ \pi\text{-}\arcsin\left[\dfrac{x(t)}{A_2}\right], & t \in [t_3, t_4) \\ 2\pi + \sin\left[\dfrac{x(t)}{A_2}\right], & t \in [t_4, t_5). \end{cases} \tag{20}$$

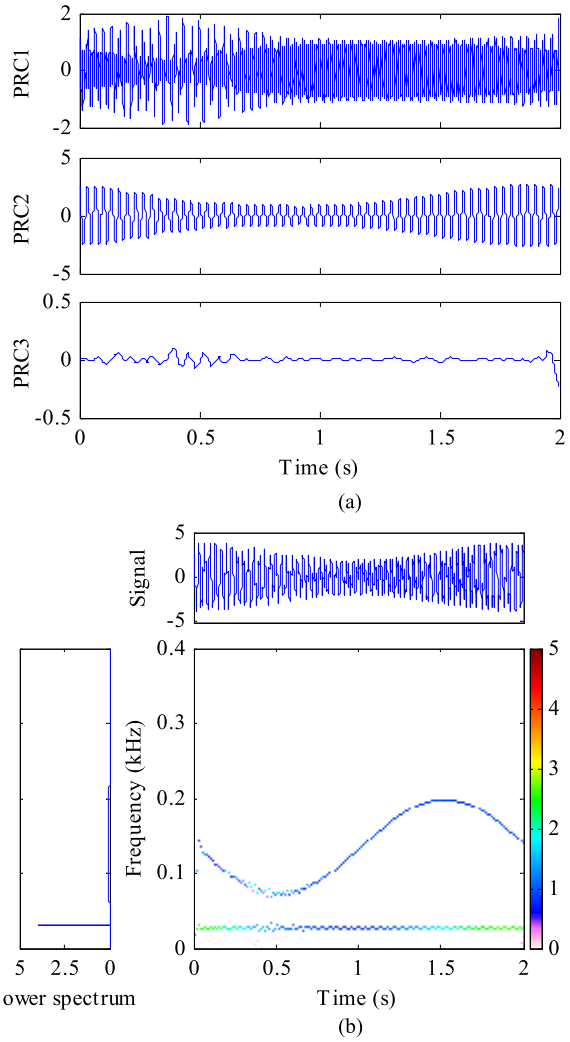

FIGURE 5. ITD analysis results. (a) PRCs; (b) Time-frequency distribution.

In order to avoid the arcsin function, and improve the computational efficiency, the instantaneous phase can be estimated via linear interpolation to the phase values at the extrema and zero crossings

$$\varphi(t) = \begin{cases} \dfrac{x(t)}{A_1}\dfrac{\pi}{2}, & t \in [t_1, t_2) \\ \dfrac{x(t)}{A_1}\dfrac{\pi}{2} + \left[1 - \dfrac{x(t)}{A_1}\right]\pi, & t \in [t_2, t_3) \\ -\dfrac{x(t)}{A_2}\dfrac{3\pi}{2} + \left[1 + \dfrac{x(t)}{A_2}\right]\pi, & t \in [t_3, t_4) \\ -\dfrac{x(t)}{A_2}\dfrac{3\pi}{2} + \left[1 + \dfrac{x(t)}{A_2}\right]2\pi, & t \in [t_4, t_5). \end{cases} \tag{21}$$

Given the instantaneous phase, the instantaneous frequency can be calculated as the time derivative of the instantaneous phase.

2) ILLUSTRATION

Fig. 5 shows the ITD analysis results. The first two PRCs correspond to the sinusoidal FM and AM components respectively. Other PRCs have small magnitudes, thus being negligible. When the instantaneous frequency of the FM component approaches the constant carrier frequency of the sinusoidal

AM component, around 0.5 s, both of them show distortion in their time–frequency distributions, see Fig. 5 (b). The time–frequency distribution resolves the frequency contents and their time-varying profiles. However, it shows small ripples around the true instantaneous frequency. In particular, when the two instantaneous frequency trajectories are close at 0.5 s, severe distortion happens to the waveform of PRC1. This is possibly due to the low time resolution of the instantaneous amplitude and instantaneous frequency estimation approach in ITD.

3) APPLICATION REVIEW

ITD has been tested in the field of machinery fault diagnosis. An and Jiang [62], [63] used ITD to decompose bearing vibration signal into PRCs, and selected the component with dominant energy to construct feature vector for fault pattern identification. Hu *et al.* [64] combined ensemble ITD, wavelet packet transform and correlation dimension for wind turbine fixed-shaft gearbox fault diagnosis. These studies demonstrate that ITD has potential in analyzing complicated and multi-component machinery vibration signals.

4) REMARKS

ITD separates the PRC in a frequency order from high to low. It does not use spline to fit local extrema. As such, it is free from the overshoot and undershoot errors in spline interpolation to the signal envelope, works better in suppressing the end effect and mode mixing, and thereby avoids possible generation of spurious extrema and shift in or exaggeration of the existing ones. Moreover, it does not involve complicated sifting process in PRC separation, thus having a low computational complexity, and avoids the smoothing of transients and time–scale smearing due to repetitive sifting. In terms of instantaneous parameter estimation, it defines the instantaneous amplitude and instantaneous frequency of PRC based on single wave analysis, thus being able to overcome the drawbacks of traditional Hilbert transform based methods, such as end effects, spikes in instantaneous frequency and negative frequency values. However, ITD uses linear transform to define the local mean baseline. This may lead to distortion in the obtained components. In addition, the instantaneous parameters are derived from single wave based on local extrema, and they are constant between any two adjacent extrema. Therefore, the time resolution is limited to the interval between consecutive extrema.

D. LOCAL CHARACTERISTIC SCALE DECOMPOSITION

1) PRINCIPLE

For a mono-component waveform, if we connect the local maxima and minima with lines respectively, then the upper and lower lines are almost symmetric about the instantaneous mean of the waveform. Inspired by this idea and the baseline extraction approach in ITD, Cheng *et al.* [15] recently proposed the local characteristic scale decomposition (LCD) algorithm. Supposes any complicated signal consists of several different intrinsic scale components (ISCs), and any two ISCs are independent of each other. To guarantee the instantaneous frequency be physically meaningful, based on the local characteristic scale parameters of extrema, the ISC is defined to meet the following two conditions:

Condition 1: In the whole data set, all the local maxima are positive, all the local minima are negative, and the signal is monotonic between any two adjacent extrema.

Condition 2: Among the whole data, denote the extrema X_k, and the corresponding occurrence time instant τ_k, where $k = 0, 1, \cdots, K$ and $K + 1$ is the number of extrema. Connecting adjacent maxima (minima) (τ_k, X_k) and (τ_{k+2}, X_{k+2}), yields a line

$$l_k(t) = (X_{k+2} - X_k)\frac{t - \tau_k}{\tau_{k+2} - \tau_k} + X_k. \tag{22}$$

Suppose the line magnitude A_{k+1}, at the occurrence time τ_{k+1} of minimum (maximum) X_{k+1}. Then, the ratio of mirror magnitude A_{k+1} to X_{k+1} is constant, i.e. $aA_{k+1}+(1-a)X_{k+1} = 0$, by default $a = 0.5$.

Condition 1 eliminates riding waves and guarantees the ISC be mono-component. Condition 2 ensures the smoothness and symmetry of the ISC waveform about the local median. These two conditions make the ISC component a single mode between two adjacent extrema and be a sinusoidal curve locally, thus guaranteeing its instantaneous frequency be physically meaningful.

Following the definition of ISC, a signal $x(t)$ can be decomposed into several mono-component ISCs through LCD method:

Step 1: Find the local extrema (including both local minima and local maxima) of $x(t)$, and denote them as (τ_k, X_k).

Step 2: Calculate the mirror magnitude

$$A_k = (X_{k+1} - X_{k-1})\frac{\tau_k - \tau_{k-1}}{\tau_{k+1} - \tau_{k-1}} + X_{k-1}, \tag{23}$$

and the center $L_k = (X_k + A_k)/2$. Note the index of A_k and L_k is $k = 2, \cdots, K - 1$. By boundary extension methods, we can obtain (τ_0, X_0) and (τ_{K+1}, X_{K+1}), and thereby L_1 and L_{K+1}.

Step 3: Construct a centerline $m(t)$ by means of cubic spline interpolation to all the center points (τ_k, L_k).

Step 4: Construct a prototype ISC $h(t) = x(t) - m(t)$.

Step 5: If $h(t)$ satisfies the ISC conditions, then set the ISC $c(t) = h(t)$. Otherwise, repeat steps 1-5 on $h(t)$.

Step 6: Construct the residual signal $r(t) = x(t) - c(t)$.

Step 7: If $r(t)$ satisfies the stop criterion for LCD, then set $r(t)$ as the final residual signal, and terminate the LCD process. Otherwise, repeat steps 1-7 on $r(t)$.

The stop criterion is defined based on the standard deviation between two consecutive sifting results for an ISC, see (4). If the standard deviation is smaller than a given threshold, for example 0.01, then stop the sifting.

LCD also separates ISCs in an order from higher frequency to lower one. Its higher computational efficiency and better performance in suppressing mode mixing and pseudo modes have been demonstrated [15].

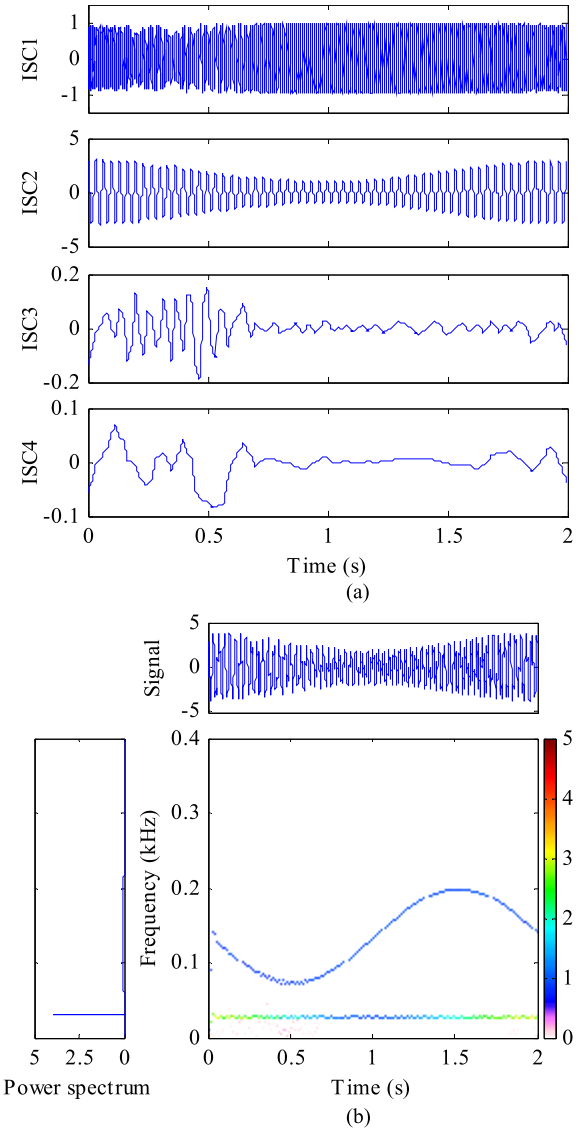

FIGURE 6. LCD analysis results. (a) ISCs; (b) Time-frequency distribution.

2) ILLUSTRATION

Fig. 6 displays the LCD analysis results. We apply the Cauchy-type stop criteria for ISC sifting, and set the standard deviation to 0.01 following the suggestion in [15]. The first two ISCs link to the sinusoidal FM and AM components respectively. In terms of waveforms, the distortion around 0.5 s, when the instantaneous frequencies of the two components are close to each other, is alleviated. The FM component has an almost constant amplitude, consistent to the theoretical expectation. The time–frequency distribution, Fig. 6 (b), has good readability, and well resolves the instantaneous frequency profile of the sinusoidal FM component and the instantaneous amplitude variability of the sinusoidal AM component.

3) APPLICATION REVIEW

A few works on application of LCD in complicated signal analysis for machinery fault diagnosis have been reported. In order to extract the intrinsic oscillations of rolling bearing fault, Zheng *et al.* [15] decomposed signals using the LCD, and extracted fault feature from the first a few ISCs. Liu *et al.* [65] used LCD to decompose signals and further extracted features for rolling bearing fault detection. These works show the possibility of LCD in effectively separating the oscillation component sensitive to fault from complicated signals.

4) REMARKS

The LCD follows the same framework as EMD. The only difference lies in the instantaneous mean extraction. It outperforms EMD in computational efficiency as well as mitigating mode mixing and pseudo modes. However, it still uses cubic spline interpolation to fit the instantaneous mean. Therefore, it also suffers from end effects and mode mixing somehow. How to address such issues deserves further investigation.

E. HILBERT VIBRATION DECOMPOSITION

1) PRINCIPLE

Inspired by the idea of EMD, Feldman [16] proposed the Hilbert vibration decomposition algorithm for time-varying complicated mechanical vibration signal analysis. HVD applies Hilbert transform and synchronous detection demodulation to estimate the instantaneous frequency and instantaneous amplitude of signals. It decomposes a complicated multi-component signal into mono-components in an order from larger instantaneous amplitude to smaller ones. This method is implemented via Hilbert transform only, but does not involve other complicated techniques, therefore its algorithm is simple [16].

HVD assumes the signal satisfies the following three conditions:

Condition 1: The signal is a superposition of several quasi-harmonic waves.

Condition 2: The instantaneous amplitude of each constituent component differs.

Condition 3: The signal duration covers more than one cycles of the slowest component.

For multi-component signals, both the instantaneous frequency estimated via Hilbert transform and the in-phase/phase shift quadrature signals include the slowly varying part corresponding to the largest amplitude and the rapidly varying part of other components. The integral of rapidly varying parts over time approaches to zero. Therefore, the instantaneous amplitude, instantaneous phase and instantaneous frequency of the slowly varying component with the largest amplitude can be estimated through low pass filter. Inspired by this idea, HVD method is designed as below:

Step 1: For a signal $x(t)$, construct its analytic signal via Hilbert transform, and estimate the corresponding instantaneous amplitude $a(t)$ and instantaneous frequency $\omega(t)$.

Step 2: Low pass filter the instantaneous frequency $\omega(t)$, yielding the instantaneous frequency $\omega_1(t)$ of the largest amplitude component.

Step 3: With the instantaneous frequency $\omega_1(t)$ as the frequency of reference signal, estimate the instantaneous amplitude $a_1(t)$ and instantaneous phase $\varphi_1(t)$ of the largest amplitude component, via synchronous detection and low pass filtering. The synchronous detection method extracts the amplitude details about a vibration component with a known frequency by multiplying the initial vibration composition by two reference signals exactly 90° out of phase with one another. The in-phase part

$$x_{l=r}(t) = \sum_l A_l(t)\cos\Big[\int \omega_l(t)dt + \varphi_l\Big]\cos\Big[\int \omega_1(t)dt\Big] = \frac{1}{2}A_l(t)\Big(\cos\varphi_l + \cos\Big\{\int [\omega_l(t)+\omega_1(t)]dt + \varphi_l\Big\}\Big), \tag{24}$$

and the phase shifted quadrature part

$$\hat{x}_{l=r}(t) = \frac{1}{2}A_l(t)\Big(\sin\varphi_l + \sin\Big\{\int [\omega_l(t)+\omega_1(t)]dt + \varphi_l\Big\}\Big), \tag{25}$$

where $A_l(t)$, $\omega_l(t)$ and φ_l are the instantaneous amplitude, frequency and phase of the lth component. Both the in-phase and the quadrature part consists of a slowly function which includes the amplitude and the phase, and a fast varying one which includes the double frequency harmonics. In such a case, it is possible to remove the oscillating part again by low pass filtration, and thus obtaining the instantaneous amplitude and instantaneous phase

$$\langle x_{l=r}(t)\rangle = \begin{cases} \frac{1}{2}A_l(t)\cos\varphi_l, & \omega_l = \omega_1 \\ 0, & \text{other,} \end{cases}$$

$$\langle \hat{x}_{l=r}(t)\rangle = \begin{cases} \frac{1}{2}A_l(t)\sin\varphi_l, & \omega_l = \omega_1 \\ 0, & \text{other,} \end{cases}, \tag{26a}$$

$$A_l(t) = 2\sqrt{\langle x_{l=r}(t)\rangle^2 + \langle \hat{x}_{l=r}(t)\rangle^2},$$

$$\varphi_l = \arctan\frac{\langle \hat{x}_{l=r}(t)\rangle}{\langle x_{l=r}(t)\rangle}. \tag{26b}$$

Then, reconstruct the largest amplitude component as $x_1(t) = a_1(t)\cos[\int \omega_1(t)dt]$.

Step 4: Subtract the largest amplitude component from the original signal, yielding a new signal $x_{l-1}(t) = x(t) - x_1(t)$. Let $x(t) = x_{l-1}(t)$, and repeat steps 1-4, until the standard error between two consecutive iterations becomes less than a predefined threshold.

In each iteration, the instantaneous amplitude and instantaneous frequency are obtained from low pass filters. The cut-off frequency of low pass filter determines the frequency resolution of HVD. The smaller the cut-off frequency, the better, when the accuracy and performance of low pass filter are satisfied.

2) ILLUSTRATION

Fig. 7 shows the HVD analysis results of the synthetic signal. We set the cut-off frequency of low pass filter

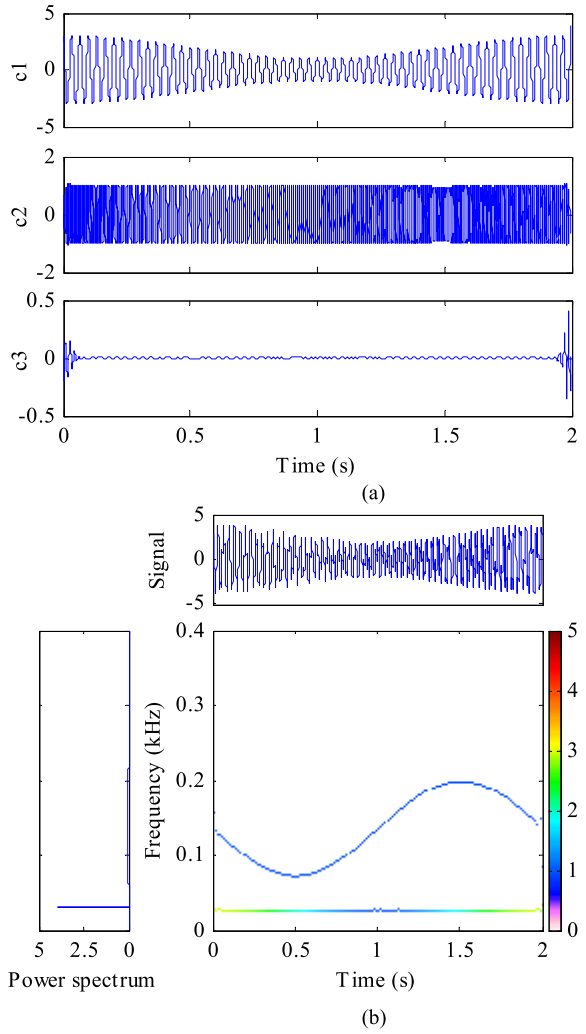

FIGURE 7. HVD analysis results. (a) Components; (b) Time-frequency distribution.

to 0.03 times the Nyquist frequency based on the recommendation in [16]. In Fig. 7 (a), the first two components correspond to the sinusoidal AM component and FM components respectively. They almost perfectly match the true components, except minor distortion near boundaries due to the end effect. The other component has minor magnitude and is negligible. The time–frequency distribution has fine resolution and good readability, see Fig. 7 (b). This helps discern the time variability in both the FM instantaneous frequency and the AM instantaneous amplitude.

3) APPLICATION REVIEW

Feldman [16]–[21] proposed the HVD for time-variant vibration signal analysis, and applied it to dynamic system identification, including moving load identification and vibration modes separation. Braun and Feldman [19] also extended the HVD to fault diagnosis of rolling bearings, gearboxes and rotors. Qin *et al.* [66] constructed time-frequency representation via HVD, and extracted the shaft misalignment

features of a gas turbine. However, wide application of HVD in machinery fault diagnosis has not been reported.

4) REMARKS

HVD is primarily suited to analysis of quasi and almost periodic signals. However, in machinery fault diagnosis, we often focus on detection of impulses and transients. Moreover, under time variant speed and loading conditions, the vibration and acoustic signals exhibit strong non-stationarity. Therefore, how to improve HVD, and extend it to transient event detection, and aperiodicity and time variability analysis, is still worth of further investigation in-depth.

F. EMPIRICAL WAVELET TRANSFORM

1) PRINCIPLE

In order to construct an adaptive signal decomposition with rigorous mathematical formulation, Gilles [22] proposed empirical wavelet transform (EWT). This algorithm builds adaptive wavelets capable of extracting AM-FM components of a signal. It is motivated by a key idea that such constituent AM-FM components have a compact support Fourier spectrum. Separating the different modes is equivalent to segment the Fourier spectrum and to apply some filtering corresponding to each detected Fourier support. The dilation factors in such wavelet do not follow a prescribed scheme such as dyadic discretization but are detected empirically according to the characteristics of signal Fourier spectrum, thus termed empirical wavelet transform.

Suppose the Fourier spectrum is separated into N continuous segments, each corresponds to a mode which centers around a specific frequency and has compact support. Assume K local maxima found in the spectrum, they are sorted in a decreasing order (0 and π are excluded). If $K \geq N$, then keep only the first $N-1$ maxima; if $K < N$, then keep all the maxima and reset N to an appropriate value.

Based on this set of maxima plus 0 and π, the boundaries of each segment are defined as the center between two consecutive maxima. Let ω_n denote the limit between each segment, where $\omega_0 = 0$ and $\omega_N = \pi$, then each segment is denoted as $\Lambda_n = [\omega_{n-1}, \omega_n]$, and the Fourier support $[0, \pi] = \cup_{n=1}^{N} \Lambda_n$.

On each segment, the empirical scaling function and empirical wavelet can be defined as band pass filters

$$\hat{\phi}_n(\omega) = \begin{cases} 1, \\ \quad |\omega| \leq \omega_n - \tau_n \\ \cos\left\{\frac{\pi}{2}\beta\left[\frac{1}{2\tau_n}(|\omega| - \omega_n + \tau_n)\right]\right\}, \\ \quad \omega_n - \tau_n \leq |\omega| \leq \omega_n + \tau_n \\ 0, \\ \quad \text{other} \end{cases} \tag{27a}$$

$$\hat{\psi}_n(\omega) = \begin{cases} 1, \\ \quad \omega_n + \tau_n \leq |\omega| \leq \omega_{n+1} - \tau_{n+1}, \\ \cos\left\{\frac{\pi}{2}\beta\left[\frac{1}{2\tau_{n+1}}(|\omega| - \omega_{n+1} + \tau_{n+1})\right]\right\}, \\ \quad \omega_{n+1} - \tau_{n+1} \leq |\omega| \leq \omega_{n+1} + \tau_{n+1} \\ \sin\left\{\frac{\pi}{2}\beta\left[\frac{1}{2\tau_n}(|\omega| - \omega_n + \tau_n)\right]\right\}, \\ \quad \omega_n - \tau_n \leq |\omega| \leq \omega_n + \tau_n \\ 0, \\ \quad \text{other} \end{cases} \tag{27b}$$

respectively, where $\tau_n = \gamma w_n$, when $\gamma < \min_n[(\omega_{n+1} - \omega_n)/(\omega_{n+1} + \omega_n)]$, the set $\{\phi_1(t), \{\psi_n(t)\}_{n=1}^{N}\}$ is a tight frame on space $L^2(R)$, $\beta(x)$ is an arbitrary $C^k([0, 1])$ function, and in this paper

$$\beta(x) = x^4(35 - 84x + 70x^2 - 20x^3). \tag{28}$$

Given the scaling and wavelet functions, EWT can be defined. The detail coefficients are given by the inner products with the empirical wavelets

$$W(n, t) = \int x(\tau)\psi_n(\tau - t)d\tau, \tag{29}$$

and the approximation coefficients (we adopt the convention to denote them) by the inner product with the scaling function

$$W(0, t) = \int x(\tau)\phi_1(\tau - t)d\tau. \tag{30}$$

The signal can be reconstructed by inverse empirical wavelet transform

$$x(t) = W(0, t) * \phi_1(t) + \sum_{n=1}^{N} W(n, t) * \psi_n(t), \tag{31}$$

where $*$ denotes convolution operator. According to this formalism, the empirical modes are defined as

$$x_0(t) = W(0, t) * \phi_1(t), \tag{32a}$$

$$x_k(t) = W(k, t) * \psi_k(t). \tag{32b}$$

Similar to wavelet transform, the EWT separate the empirical mode in a frequency order from low to high, but the bandwidth is not dyadic since the frequency band is segmented empirically.

The number of modes to be separated is a key parameter in EWT. Gilles [22] proposed an empirical method to determine the number. Let $\{M_k\}_{k=1}^{K}$ denote the set of K local maxima in the signal Fourier spectrum. Assume this set is sorted in a decreasing order ($M_1 \geq M_2 \geq \cdots M_K$) and normalized in $[0, 1]$. Usually, the most important maxima are significantly larger than the other maxima, i.e. the meaningful maxima are greater than some amount of the difference between the biggest maximum M_1 and the smallest maximum M_K. In practice, we can keep all maxima larger than a predefined threshold $M_K + \alpha(M_1 - M_K)$, where the relative magnitude ratio α around 0.3 and 0.4, which corresponds to a tradeoff between too much detection and a good separation of the information in the Fourier spectrum.

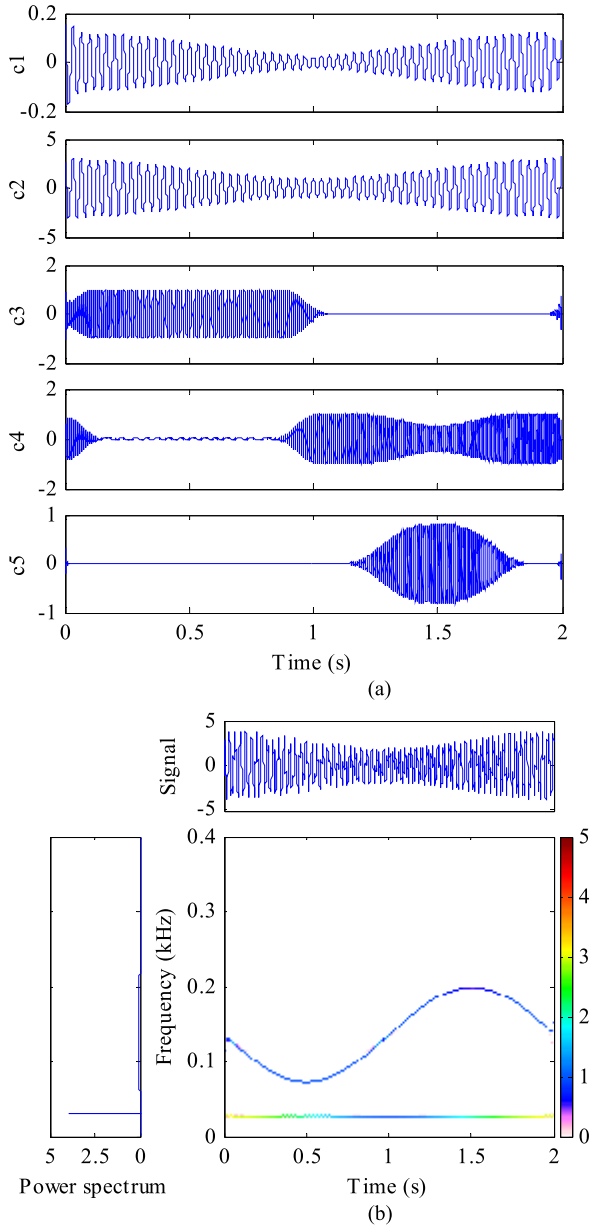

FIGURE 8. EWT analysis results. (a) Components; (b) Time-frequency distribution.

2) ILLUSTRATION

Fig. 8 shows the EWT analysis results of the synthetic signal. Following the recommended spectrum segmentation method in [22], the spectrum is divided into five subbands, therefore five components are obtained, as shown in Fig. 8 (a). Unfortunately, the EWT result does not preserve the integrity of each constituent component, and they are split in frequency domain. The first two together correspond to the sinusoidal AM component, and the third through fifth together link to the sinusoidal FM component. Anyway, the time-frequency distribution of raw signal is a superposition of all the empirical modes' time-frequency representation, thus may alleviate the spectral splitting effect. The time-frequency distribution, Fig. 8 (b), recovers the frequency contents and their time variability almost perfectly, even though some distortions exist at the instants when the constituent components are split around 0.1 s, 0.9 s and 1.5 s, and small ripples emerge when the two instantaneous frequency trajectories are close.

3) APPLICATION REVIEW

EWT has also been tried in signature extraction for machinery fault diagnosis. Kedadouche *et al.* [67], [68] defined the support boundaries of empirical wavelet filters based on operational modal analysis, proposed a kurtosis based index for selecting sensitive empirical modes, and detected rolling bearing fault via envelope spectrum analysis of selected empirical modes. Pan *et al.* [69] used EWT to separate impulsive components, and thereby detected wind turbine generator bearing fault. Chen *et al.* [70] proposed an adaptive spectrum segmentation approach based on the local minima of scale representation for EWT, and extracted the inherent modulation feature of rolling bearing faults. Cao *et al.* [71] detected the transient impulses by EWT, and thereby diagnosed train wheel bearing faults. In order for decoupling diagnosis of rolling bearing compound fault, Jiang *et al.* [72] proposed an approach based on EWT and Duffing oscillator. They established the fault isolator by incorporating each single fault frequency with Duffing oscillator, separated fault signature into empirical modes via EWT, fed the empirical modes into the fault isolator, and identified each single fault by studying the chaotic motion in Poincare mapping of fault isolator outputs. These works have demonstrated the attractive capabilities of EWT in machinery fault feature extraction.

4) REMARKS

EWT differs from conventional wavelet transform mainly in non-dyadic partition of frequency band. How to segment signal spectrum is necessary to well extract the constituent empirical modes, and still needs further investigation in-depth. This involves the number of modes to be seprated and the frequency boundaries of each mode. To accurately determine the number of modes, it could be interesting to use the concept of best basis in the wavelet packets transform. Usually, the center between two adjacent local maxima is taken as frequency boundaries, but this does not use the information of the spectrum shape and returns "perturbed" modes. A clustering viewpoint could provide a solution to both problems.

G. VARIATIONAL MODE DECOMPOSITION

1) PRINCIPLE

Recently, Dragomiretskiy and Zosso [23] proposed the variational mode decomposition (VMD). This algorithm can non-recursively decompose a complicated multi-component signal into constituent AM-FM components, and it is robust to noise. VMD is an entirely non-recursive decomposition method, and it extracts the constituent AM-FM components of a complicated multi-component signal adaptively and

concurrently. It defines IMFs as explicit AM-FM models, and relates the parameters of AM-FM models to the bandwidth of IMFs. According to the narrow-band property of IMFs, the AM-FM parameters can be found by minimizing the bandwidth, thus obtaining IMFs. This algorithm has good merits over other available mode decomposition methods, such as theoretical rationale and robustness to noise and sampling.

In essence, IMFs are AM-FM signals, and they have a limited bandwidth. VMD decomposes a signal $x(t)$ into an ensemble of IMFs $c_k(t)$ that are band-limited about their respective center frequency ω_k, while reconstructing the signal optimally. It iteratively updates each IMF $c_k(t)$ in the frequency domain, and then estimates the center frequency ω_k as the center of gravity of the IMF power spectrum.

Motivated by the narrow-band properties of the AM-FM IMF definition, each IMF $c_k(t)$ is assumed to be mostly compact around a center frequency ω_k, i.e. it has specific sparsity properties. The sparsity prior of each IMF is described by its bandwidth. For each IMF $c_k(t)$, in order to assess its bandwidth, its analytic signal is firstly computed by means of Hilbert transform to obtain a spectrum of unilateral non-negative frequency. Then, its spectrum is shifted to baseband by multiplying with an exponential harmonic tuned to the respective center frequency. Finally, the bandwidth can be estimated through the squared l_2 norm of the gradient. The resulting constrained variational optimization problem is

$$\begin{aligned}&\min_{\{c_k(t)\},\{\omega_k\}} \quad \sum_k \left\| \frac{\partial}{\partial t}\left\{\left[\delta(t)+j\frac{1}{\pi t}\right]* c_k(t)\exp\right\}(-j\omega_k t)\right\|_2^2,\\ &\text{s.t.} \quad \sum_{k=1}^{K} c_k(t) = x(t),\end{aligned} \tag{33}$$

where $\delta(\cdot)$ is the Dirac delta function, the symbol $*$ denotes the convolution operator, and K is the number of IMFs to be extracted.

In order to render the optimization problem, (33), into an unconstrained form, a quadratic penalty term and a Lagrangian multiplier are introduced, for quick convergence and strict enforcement of the constraint. Then the objective function to be minimized becomes an augmented Lagrangian

$$\begin{aligned}&L[\{c_k(t)\},\{\omega_k\},\lambda(t)]\\ &= \alpha\sum_k \left\| \frac{\partial}{\partial t}\left\{\left[\delta(t)+j\frac{1}{\pi t}\right]* c_k(t)\exp\right\}(-j\omega_k t)\right\|_2^2\\ &\quad + \left\| x(t)-\sum_k c_k(t)\right\|_2^2 + \left\langle \lambda(t), x(t)-\sum_k c_k(t)\right\rangle,\end{aligned} \tag{34}$$

where $\lambda(t)$ is the Lagrange multiplier, α is the balancing parameter of the data-fidelity constraint, and $\langle\cdot,\cdot\rangle$ stands for inner product.

The solution to the minimization problem, (33) can now be found as the saddle point of the augmented Lagrangian (34), in a sequence of iterative sub-optimizations [23].

Each IMF $c_k(t)$ can be updated as a solution to a minimization problem equivalent to (34)

$$\begin{aligned}&c_k(t)\\ &= \arg\min_{c_k} L(\{c_k(t)\},\{\omega_k\},\lambda(t)]\\ &= \arg\min_{c_k}\left(\alpha\sum_k \left\| \frac{\partial}{\partial t}\left\{\left[\delta(t)+j\frac{1}{\pi t}\right]* c_k(t)\exp\right\}(-j\omega_k t)\right\|_2^2\right.\\ &\quad \left. + \left\| x(t)-\sum_k c_k(t)+\frac{\lambda(t)}{2}\right\|_2^2\right).\end{aligned} \tag{35}$$

In the frequency domain, the solution to (35) can be found as [23]

$$\hat{c}_k(\omega) = \frac{\hat{x}(\omega)-\sum_{i\neq k}\hat{c}_i(\omega)+\frac{1}{2}\hat{\lambda}(\omega)}{1+2\alpha(\omega-\omega_k)^2}. \tag{36}$$

Then, the IMF in the time domain $c_k(t)$ can be obtained by inverse Fourier transforming (36) and taking the real part.

The center frequency ω_k associated with each IMF $c_k(t)$ can also be updated as a solution to a minimization problem equivalent to (34)

$$\begin{aligned}\omega_k &= \arg\min_{\omega_k} L(\{c_k(t)\},\{\omega_k\},\lambda(t)]\\ &= \arg\min_{\omega_k}\sum_k \left\| \frac{\partial}{\partial t}\left\{\left[\delta(t)+j\frac{1}{\pi t}\right]* c_k(t)\exp\right\}(-j\omega_k t)\right\|_2^2.\end{aligned} \tag{37}$$

It can be found as the center of gravity of the associated IMF's power spectrum [23]

$$\omega_k = \frac{\int_0^\infty \omega\left|\hat{c}_k(\omega)\right|^2 d\omega}{\int_0^\infty \left|\hat{c}_k(\omega)\right|^2 d\omega}. \tag{38}$$

The complete algorithm of VMD is summarized as follows:

Step 1: Initialization: Let $\{\hat{c}_k^0(t)\}$, $\{\hat{\omega}_k^0\}$, $\hat{\lambda}^0(t)$, n be 0, and predefine convergence threshold ε and the number of IMFs K to be separated.

Step 2: Update each IMF $c_k(t)$ and its associated center frequency ω_k, for $k = 1:K$ and all $\omega \geq 0$

$$\hat{c}_k^{n+1}(\omega) = \frac{\hat{x}(\omega)-\sum_{i<k}\hat{c}_i^{n+1}(\omega)-\sum_{i>k}\hat{c}_i^{n}(\omega)+\frac{1}{2}\hat{\lambda}^n(\omega)}{1+2\alpha(\omega-\omega_k^n)^2}, \tag{39}$$

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega\left|\hat{c}_k^{n+1}(\omega)\right|^2 d\omega}{\int_0^\infty \left|\hat{c}_k^{n+1}(\omega)\right|^2 d\omega}. \tag{40}$$

Step 3: Update the Lagrangian multiplier, for all $\omega \geq 0$

$$\hat{\lambda}^{n+1}(\omega) = \hat{\lambda}^n(\omega)+\tau\left[\hat{x}(\omega)-\sum_k \hat{c}_k^{n+1}(\omega)\right], \tag{41}$$

where τ is the Lagrangian multiplier update parameter.

*Step 4:*Check the convergence condition

$$\sum_k \frac{\left\|\hat{c}_k^{n+1}(t)-\hat{c}_k^n(t)\right\|_2^2}{\left\|\hat{c}_k^n(t)\right\|_2^2} < \varepsilon. \tag{42}$$

If it is met, let $c_k(t)=\hat{c}_k^{n+1}(t)$ and $\omega_k=\omega_k^{n+1}$, and terminate the decomposition. Otherwise, let $n=n+1$, return to step 2.

VMD decomposes a complicated signal into a specific number of IMFs. These IMFs are AM-FM in nature, thus we can estimate their instantaneous amplitude and instantaneous frequency. More importantly, the non-recursive and concurrent decomposition nature of VMD effectively avoids the shortcoming of recursive decomposition algorithms, such as sensitivity to noise and sampling, over-shooting or under-shooting of upper and lower envelope via interpolation to extrema.

2) ILLUSTRATION

Fig. 9 shows the VMD analysis results of the synthetic signal. We set the number of IMFs and the tolerance to 4 and 1E-5 respectively, following the suggestions in [23]. The sinusoidal AM component is well separated to be IMF1, as shown in Fig. 9 (a), but the sinusoidal FM component is split and distributed over IMF2-4. This defect is also caused by the spectrum splitting effect, and harms the integrity of sinusoidal FM component, but it can be alleviated in the time-frequency distribution by superposing all IMFs' time-frequency representation. Fig. 9 (b) recovers the time-frequency structure of the synthetic signal, despite some small ripples around the true instantaneous frequencies.

3) APPLICATION REVIEW

Wang and his colleagues [73]–[75] discovered the wavelet packet like frequency band decomposition property of VMD based on fractal Gaussian noise simulations, and applied VMD to extract the fundamental, sub-harmonics, super-harmonics, and impacts of a gas turbine rotor-stator rubbing fault, as well as impulsive components of rolling bearing fault. An and Zeng [76] utilized VMD to analyze the nonstationary pressure fluctuation signal of a hydraulic turbine draft tube. They showed that VMD is better than EMD in suppressing mode mixing and improving time-frequency readability. Tang *et al.* [77] proposed to decompose signals into multiple components, and then use independent component analysis to solve the underdetermined blind source separation problem. They detected the compound fault of rolling bearings with the proposed method. Lv *et al.* [78] extracted features from the VMD based time-frequency representation, and identified rolling bearing faults. Yi *et al.* [79] improved the robustness of VMD to sampling and noise via particle swarm optimization, and utilized the proposed method to detect rolling bearing faults. Mahgoun *et al.* [80] used VMD to detect defect impulses under variable speeds and loads, and verified the feasibility via analysis of gear transmission dynamics simulated signals. These studies have demonstrated that VMD can

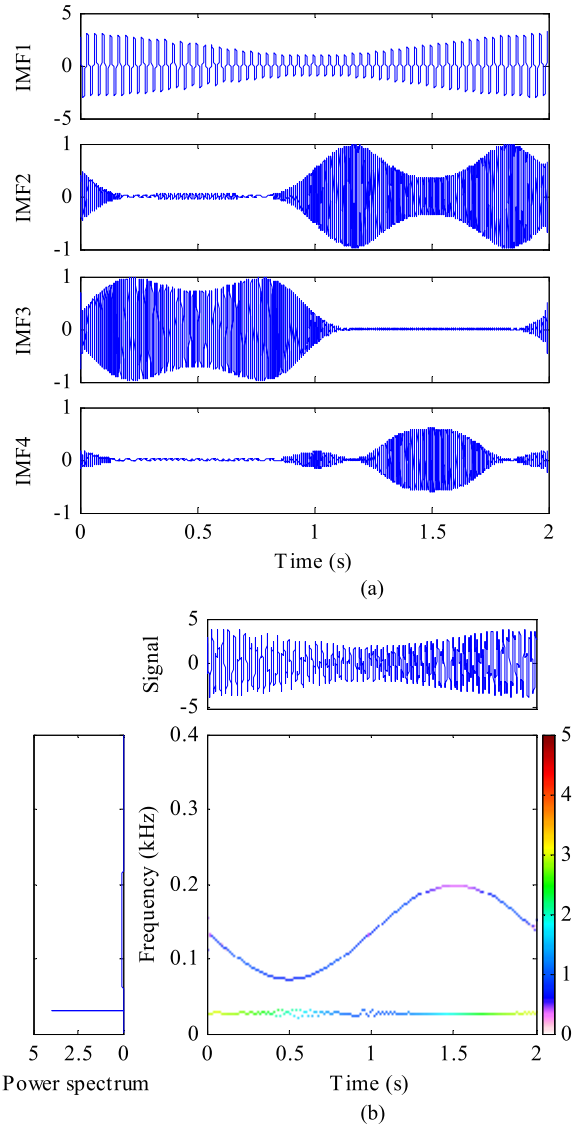

FIGURE 9. VMD analysis results. (a) IMFs; (b) Time-frequency distribution.

effectively decompose complicated multi-component signals into AM-FM components, and suppress noise interferences, thus enhancing fault features.

4) REMARKS

VMD shares the same nature as EWT, because they are all based on spectrum segmentation, but VMD goes further since it considers the spectral shape and takes the gravity center of spectrum as the center frequency of each IMF. However, how to determine the number of IMFs to be separated is a key issue in VMD.

H. NONLINEAR MODE DECOMPOSITION PRINCIPLE

1) PRINCIPLE

More recently, Iatsenko *et al.* [24] proposed the nonlinear mode decomposition (NMD) for nonlinear and nonstationary signal analysis. NMD decomposes an arbitrary complicated signal into a set of physically meaningful modes for any waveform, and in the meantime removes noise. It is designed

based on a combination of time-frequency analysis, surrogate data tests and harmonic identification, thus being robust to noise and able to identify interdependent oscillations and distinguish deterministic from random activity. "Nonlinear" in NMD is irrelevant to classical nonlinear analysis, but comes from the fact that the resultant modes from the decomposition method have complicated waveforms which are commonly due to nonlinearities in either the generating system or the measurement.

Signals generated from complicated systems are usually composed of a mixture of different oscillations. They are rarely purely sinusoidal, but show more complicated wave shapes due to nonlinearities in the generating system and/or the measurement apparatus, and are often characterized by time-varying amplitudes and frequencies. To better consider the complexity of such wave shape, nonlinear modes (NMs) are defined as a sum of all AM-FM components corresponding to the same activity,

$$c(t) = A(t)h[\phi(t)] = A(t)\sum_i a_i \cos[i\phi(t) + \varphi_i], \quad (43)$$

where the wave shape function $h[\phi(t)] = h[\phi(t) + 2\pi]$ is a periodic function of phase, which can be expanded as a Fourier series due to its periodicity.

An arbitrary signal is assumed to consist of several NMs $c_m(t)$ plus some noise $n(t)$

$$x(t) = \sum_m c_m(t) + n(t). \quad (44)$$

The ultimate goal of NMD is to extract all the constituent NMs, and to find their characteristic parameters including amplitudes $A(t)$, phases $\phi(t)$, and frequencies $f(t)$, as well as the amplitude scaling factors a_i and phase shifts φ_i of the harmonics. The NMD algorithm is summarized as follows.

Step 1: Calculate the time-frequency representation of the given signal based on the log-normal wavelet

$$\hat{\psi}(\omega) = \exp[(-\pi f_0 \ln \omega)^2], \quad \omega_\psi = 1, \quad (45)$$

where f_0 is the resolution parameter determining the trade-off between time and frequency resolution (by default, it is set to $f_0 = 1$).

Step 2: Extract the dominant component (reference component) from the time–frequency representation, and reconstruct its characteristic parameters (including instantaneous amplitude, instantaneous phase and instantaneous frequency) through direct or ridge method Direct:

$$A(t)\exp[j\phi(t)] = C_\psi^{-1}\int_{\omega_-(t)}^{\omega_+(t)} \mathrm{WT}(t,\omega)\frac{d\omega}{\omega},$$
$$C_\psi = \frac{1}{2}\int_0^\infty \hat{\psi}(\omega)\frac{d\omega}{\omega}, \quad (46a)$$
$$\omega(t) = Re\left[\frac{D_\psi^{-1}\int_{\omega_-(t)}^{\omega_+(t)} \omega \mathrm{WT}(t,\omega)\frac{d\omega}{\omega}}{C_\psi^{-1}\int_{\omega_-(t)}^{\omega_+(t)} \mathrm{WT}(t,\omega)\frac{d\omega}{\omega}}\right],$$
$$D_\psi = \frac{\omega_\psi}{2}\int_0^\infty \frac{1}{\omega}\hat{\psi}^*(\omega)\frac{d\omega}{\omega}, \quad (46b)$$

where $[\omega_-(t), \omega_+(t)]$ is estimated as the widest region of unimodal and nonzero time–frequency representation amplitude around the time–frequency ridge $\omega_p(t) = \arg\max_\omega |\mathrm{TFR}(t,\omega)|$ at each time t, and $*$ denotes complex conjugate.

Ridge:

$$\omega(t) = \omega_p(t)\exp\{\delta \ln[\omega_d(t)]\}, \quad (47a)$$
$$A(t)\exp[j\phi(t)] = \frac{2\mathrm{WT}[t,\omega_p(t)]}{\hat{\psi}^*[\omega_\psi \omega(t)/\omega_p(t)]}, \quad (47b)$$

where $\omega_p(t)$ is the time–frequency ridge curve, and $\delta \ln[\omega_d(t)]$ is the correction for discretization effects found by parabolic interpolation.

Usually, the direct method better recovers the time variations of amplitude and frequency but is less robust to noise and interferences than the ridge method.

Step 3: Test the reference component against noise using surrogates test method. The surrogates test criterion is motivated by the following idea: if a component is true (and not just formed from noise peaks picked in the time–frequency plane), then it is expected to have more deterministic amplitude modulation and frequency modulation than the surrogate components, which should be more stochastic; otherwise, there will be no difference.

Construct the surrogates via Fourier transform, by inverse Fourier transforming the Fourier transform of the component with randomized phases of the Fourier coefficients. The degree of order can be quantified by spectral entropy. So the discriminating statistics for the surrogate test can be taken as a combination of the spectral entropies of the extracted amplitude $A(t)$ and frequency $f(t)$

$$D(\alpha_A, \alpha_f) = \alpha_A Q[\hat{A}(\omega)] + \alpha_f Q[\hat{f}(\omega)], \quad (48)$$

where the spectral entropy

$$Q[h(x)] = -\int \frac{|h(x)|^2}{\int |h(x)|^2 dx} \ln \frac{|h(x)|^2}{\int |h(x)|^2 dx} dx. \quad (49)$$

Calculate $D(1,0)$, $D(0,1)$ and $D(1,1)$, and select the significance as the maximum among them.

By default, generate $N_s = 40$ surrogates and set a significance level to $\lambda = 95\%$, rejecting the tested null hypothesis of noise if the number of surrogates with $D_s > D_0$ (where D_0 is the significance of the original component) is equal or higher than $N_s \times \lambda = 0.95 \times 40 = 38$.

If at least for one of them the null hypothesis is rejected, we regard the component as true one but not noise, and continue the decomposition. Stop the decomposition if it does not pass this test.

Step 4: Check whether the subharmonics of the extracted reference component is the fundamental harmonic. For $i = \frac{1}{2}, \frac{1}{3},$, do the following.

4.1 To alleviate the computational burden, calculate the signal time-frequency representation within the time-frequency support

$$\omega_\mp^{(i)} = i\left\langle \omega_p^{(1)}(t)\right\rangle + \max(1, i)[\omega_\mp^{(1)} - \left\langle \omega_p^{(1)}(t)\right\rangle], \quad (50)$$

using different resolution parameter $f_0^{(i)}$,

$$f_0^{(i)} = \begin{cases} \frac{1}{i} f_0^{(1)} \min(1, i), & \text{for STFT} \\ f_0^{(1)} \min(1, i), & \text{for WT,} \end{cases} \tag{51}$$

for each of which:

(i) Extract the ith harmonic of the reference component from the time–frequency representation, and reconstruct its amplitude, phase, and frequency. Given the fundamental frequency $\omega_p^{(1)}(t)$, the time–frequency ridge of ith harmonic $\omega_p^{(i)}(t)$ is expected to lie in the same time–frequency support as $i\omega_p^{(1)}(t)$. According to equations in step 2, the parameters of ith harmonic can be reconstructed.

(ii) Check whether the current harmonic is a true one, using surrogate data method to test against the null hypothesis of independence between the first harmonic and the extracted harmonic candidate. The following measures quantify the degree of consistency between the first harmonic and the extracted harmonic candidate, in terms of amplitude, phase and frequency respectively

$$q_A^{(i)} = \exp\left\{ - \frac{\sqrt{\left\langle [A^{(i)}(t)\left\langle A^{(1)}(t)\right\rangle - [A^{(1)}(t)\left\langle A^{(i)}(t)\right\rangle]^2 \right\rangle}}{\left\langle A^{(1)}(t) A^{(i)}(t)\right\rangle} \right\}, \tag{52a}$$

$$q_\phi^{(i)} = a \left| \left\langle \exp\left\{ j[\phi^{(i)}(t) - i\phi^{(1)}(t)] \right\} \right\rangle \right|, \tag{52b}$$

$$q_f^{(i)} = \exp\left\{ - \frac{\sqrt{\left\langle [f^{(i)}(t) - if^{(1)}(t)]^2 \right\rangle}}{\left\langle f^{(i)}(t)\right\rangle} \right\}. \tag{52c}$$

An overall measure of interdependence between the harmonics is constructed as

$$\rho^{(i)}(\beta_A, \beta_\phi, \beta_f) = [q_A^{(i)}]^{\beta_A} [q_\phi^{(i)}]^{\beta_\phi} [q_f^{(i)}]^{\beta_f}, \tag{53}$$

where parameters $\beta_{A,\phi,f}$ defines weights to each of the consistencies $q_{A,\phi,f}^{(i)}$, and $\rho^{(i)} = \rho^{(i)}(1, 1, 0)$ by default.

To eliminate the noise interference on the consistency, we employ the idea of time-shifted surrogate, and calculate the consistency $\rho_{d=1,\cdots,N_d}^{(i)}(1, 1, 0)$ from time-shifted time–frequency representations.

The probability for the extracted ith harmonic curve being a true harmonic of the main one is quantified by the significance of the surrogate test, i.e., by the ratio of number of surrogates for which $\rho_d^{(i)} < \rho_0^{(i)}$ to the total number of surrogates. A harmonic is regarded as true if the probability is equal to or greater than 95%.

To eliminate the pseudo harmonics, a threshold $\rho_{\min} = 0.5^{(\beta_A+\beta_\phi)}$ (by default $\rho_{\min} = 0.25$) is imposed on the probability, i.e. $\rho^{(i)} \geq \rho_{\min}$.

Then, a harmonic regarded as true if and only if it both passes the surrogate test and satisfies the threshold condition.

4.2 If for some $f_0^{(i)}$, the harmonic is true, then set its characteristic parameters to those reconstructed for the $f_0^{(i)}$ that is characterized by the highest consistency (52) with the reference component among $f_0^{(i)}$ for which the harmonic is identified as true.

4.3 Stop when a predefined number (default 3) of consequent harmonics are identified as false for all tested $f_0^{(i)}$.

Step 5: If some harmonic is identified as true, take the true harmonic with the smallest i as the reference component, which is guaranteed to be the first harmonic of the corresponding NM.

Step 6: Perform step4 for $i = 2, 3, \cdots$, and store the reconstructed parameters of the harmonics.

Step 7: Based on the parameters of the true harmonics, reconstruct the full NM. The parameters of each harmonic are refined by weighted averaging over the parameters of all harmonics

$$\hat{A}^{(i)}(t) = \left\langle A^{(i)}(t)\right\rangle \frac{\sum_{i'} A^{(i')}(t)}{\sum_{i'} \left\langle A^{(i')}(t)\right\rangle}, \tag{54a}$$

$$\hat{\phi}^{(i)}(t) = \arg\left(\sum_{i'} \min\left(1, \frac{i'}{i}\right) \left\langle A^{(i')}(t)\right\rangle \exp\left[j \frac{i\phi^{(i')}(t) - \Delta\phi_{i',i}}{i'} \right] \times \exp\left\{ -j \frac{2\pi}{i'} \text{round}\left[\frac{i\phi^{(i')}(t) - i'\phi^{(i)}(t) - \Delta\phi_{i',i}}{2\pi} \right] \right\} \right), \tag{54b}$$

$$\hat{\omega}^{(i)}(t) = \frac{\sum_{i'} \min(1, \frac{i'}{i}) A^{(i')}(t) i\omega^{(i')}/i'}{\sum_{i'} \min(1, \frac{i'}{i}) \left\langle A^{(i')}(t)\right\rangle}, \tag{54c}$$

where $\Delta\phi_{i',i} = \arg(\exp\{j[i\phi^{(i')}(t) - i'\phi^{(i)}(t)]\})$.

Step 8: Subtract the reconstructed NM from the signal, and repeat steps 1-7 on the residual.

Through NMD method, the full underlying oscillation modes of any wave form can be recovered. It is highly adaptive, because most of its settings are automatically adapted to the signal characteristics. Meanwhile, it is robust to noise, and outputs physically meaningful modes, with a residue as noise.

2) ILLUSTRATION

Fig. 10 shows the NMD analysis results. We utilize direct method to estimate instantaneous parameters, and set all NMD parameters to default values recommended in [24]. The first two NMs exactly represent the sinusoidal AM and FM components respectively, and their waveforms perfectly reflect the temporal behavior of the true components, as shown in Fig. 10 (a). Even though a third NM is also extracted, its magnitude is so small that can be neglected. The fine mono-component decomposition of NMD helps well reveal the feature of each component, and construct a high quality time–frequency representation, as shown in Fig. 10 (b). The time–frequency distribution has a high time–frequency resolution, is free from both inner and outer interferences, thus featuring a better readability. These merits benefit identification of the frequency contents and their time variability. As such, the time–frequency structure of the two components is well extracted, exactly consistent with the true theoretical settings.

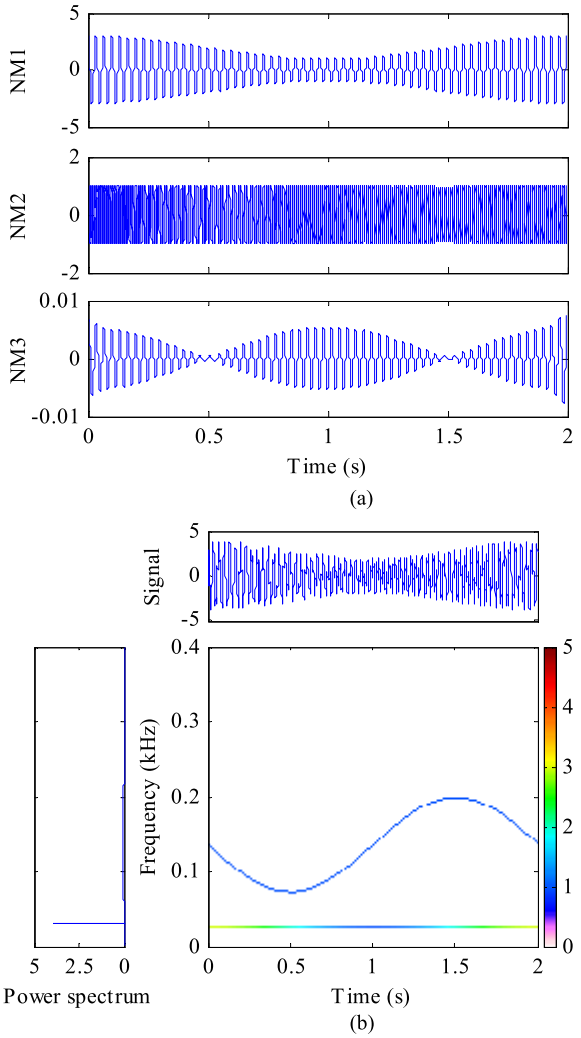

FIGURE 10. NMD analysis results. (a) NMs; (b) Time-frequency distribution.

3) APPLICATION REVIEW

Although NMD has excellent merits, it has not been widely applied in machinery fault diagnosis. Only a few relevant researches have been reported so far. Clemson *et al.* [81] highlighted the advantages of NMD in complicated oscillatory mode separation, and suggested it for characterizing the underlying time-dependent dynamics of living systems. Zhang *et al.* [82] exploited the noise resistance and mono-component decomposition capabilities of NMD to improve the performance of adaptive optimal kernel time–frequency representation, and applied it to extraction of time–frequency structure of underwater acoustic signals.

4) REMARKS

Although NMD has good merits such as robustness to noise and fine recovery of each individual complicated oscillation mode, it suffers from high computational complexity due to the heavy burden in time–frequency ridge extraction and reconstruction, and surrogate data test as well. In addition, its performance still essentially relies on traditional time–frequency analysis. Therefore, how to alleviate the computational burden and further improve its time–frequency resolving capability through appropriate time–frequency representation methods still need investigation.

I. ADAPTIVE LOCAL ITERATIVE FILTERING

1) PRINCIPLE

In EMD, the instantaneous mean is defined as the mean function of upper and lower envelopes. It is unstable under perturbations, because cubic splines, which are susceptible to singularities, are used to fit the upper and lower envelopes by connecting local maxima and local minima respectively.

To overcome this drawback, Lin *et al.* [26] proposed an iterative filtering algorithm. It follows the same algorithm framework as EMD, but instead derives the instantaneous mean by low pass filtering the signal. To guarantee the stability under perturbation and convergence, uniform double average filter of fixed length is used.

However, to effectively analyze non-linear and non-stationary signals, filters with compact support and flexible length along time is highly desirable. To this end, Cicone *et al.* [25] proposed an adaptive local iterative filtering (ALIF) algorithm. It generalizes the existing iterative filtering algorithm using non-uniform filters.

For signal filtration, long support filters are unsuitable for transients detection, because they may mix features that are far apart in a signal and contaminates the true event. However, compact support low pass filters, such as the double average filters, are not smooth enough, resulting in pseudo oscillations in subsequent IMFs. To overcome this drawback, filters are designed as the solution to the Fokker–Planck (FP) equation,

$$\frac{\partial}{\partial t}g(x,t) = -\alpha\frac{\partial}{\partial x}[p(x,t)g(x,t)] + \beta\frac{\partial^2}{\partial x^2}[q^2(x,t)g(x,t)], \alpha, \beta > 0, \quad (55)$$

inspired by the diffusion process in partial differential equations. These FP filters have compactly support, and are infinitely differentiable and smoothly vanishing to zero at both ends. Such properties avoid the pseudo oscillations during the iterative filtering process.

To capture the changes in a non-stationary signal, the filter length must be adapted along time accordingly. Given a signal, the interval between consecutive local extrema reflects the local average period of the highest frequency component. Based on this idea, the adaptive filter length is set as a multiple of the interval between consecutive local minimum and maximum. To guarantee a smooth and continuous filter length over time, the local extrema series is interpolated, and then low pass filtered to remove high frequency oscillations.

For a signal $x(t)$, based on the above adaptive local filter design approach, ALIF algorithm is summarized as below:

Step 1: Design an adaptive local FP filter $g(t, \tau)$, and find its time-varying filter length $l(t)$.

Step 2: Calculate the instantaneous mean

$$m(t) = \int_{-l(t)}^{l(t)} x(t+\tau)g(t,\tau)d\tau. \tag{56}$$

Step 3: Construct a prototype IMF $h(t) = x(t) - m(t)$.

Step 4: If $h(t)$ satisfies the IMF conditions, then set the IMF $c(t) = h(t)$. Otherwise, repeat steps 1-4 on $h(t)$.

Step 5: Construct a residual signal $r(t) = x(t) - c(t)$.

Step 6: If $r(t)$ satisfies the stop criterion for ALIF, then set $r(t)$ as the final residual signal, and terminate the ALIF process. Otherwise, repeat steps 1-6 on $r(t)$.

The stop criterion for the sifting process in step4 can be set according to the standard deviation of two consecutive sifted results, see (4). When the standard deviation reaches a predefined threshold, the sifting process stops. The ALIF algorithm stops when the residual signal $r(t)$ becomes a trend, i.e. it has one local extremum at most.

2) ILLUSTRATION

Fig. 11 shows the ALIF decomposition analysis results. We apply the Cauchy-type stop criteria for IMF sifting, and set the standard deviation to 6E-5 following the suggestion in [25]. The first two IMFs represent the sinusoidal FM and AM components respectively, and the residue has a small magnitude so that can be neglected, as shown in Fig. 11 (a). In the time–frequency distribution, Fig. 11 (b), some ripples appear around 0.5 s for the sinusoidal FM component. Meanwhile, the instantaneous frequency of sinusoidal AM component fluctuates in the duration from 0.5 s to 1.5 s, even it has a constant carrier frequency. This is mainly caused by the error of instantaneous frequency calculation.

3) APPLICATION REVIEW

Application of the ALIF algorithm in machinery fault diagnosis or in other signal analysis relevant fields has been very limited, because it is one of the most recently proposed algorithms. An *et al.* [83] utilized ALIF to decompose rolling bearing vibration signals, and extracted fault features from the envelope spectrum of AM-FM mono-component. An *et al.* [84] further applied ALIF to multi-scale analysis of vibration signals, and constructed feature vector based on the selected IMF by singular vector decomposition for wind turbine roller bearing fault diagnosis.

4) REMARKS

The local filter is a key factor in the ALIF algorithm. The filter form is set as a solution to the FP equation, and the support length in time domain is adapted to the interval between consecutive local extrema. The cut-off frequency of low pass filter is another important parameter for the local filter to effectively separate the instantaneous mean. How to set this parameter has not been well studied.

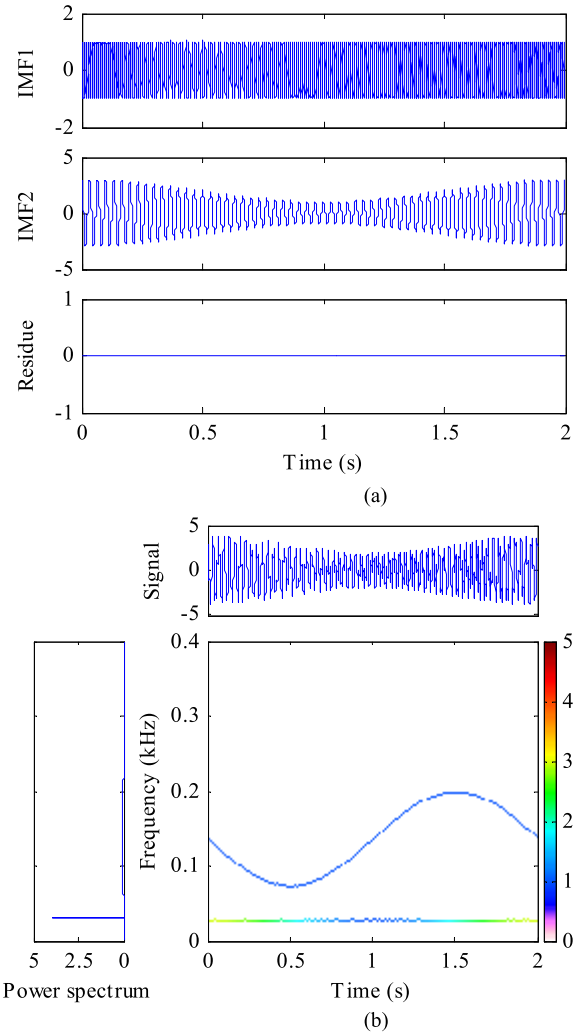

FIGURE 11. ALIF analysis results. (a) IMFs and residue; (b) Time-frequency distribution.

IV. INSTANTANEOUS FREQUENCY ESTIMATION APPROACHES

A. PRINCIPLE

Instantaneous frequency is necessary for revealing the frequency contents and their time variability of nonstationary signals, and thereby understanding the detailed generating mechanisms and the true physical nature reflected by signals. To finely estimate the instantaneous frequency of a mono-component, various approaches have been proposed, including analytic signal, direct quadrature, normalized Hilbert transform, energy separation and generalized zero-crossing approaches.

1) HILBERT TRANSFORM BASED ANALYTIC SIGNAL

The most well-known instantaneous frequency estimation approach is based on analytic signal via Hilbert transform. Given a real signal $x(t)$, its analytic signal is defined as

$$z(t) = x(t) + j\mathrm{HT}[x(t)] = a(t)\exp[j\phi(t)], \tag{57}$$

where, the Hilbert transform

$$\mathrm{HT}[x(t)] = \frac{p}{\pi}\int_{-\infty}^{\infty}\frac{x(\tau)}{t-\tau}d\tau, \tag{58}$$

p is the Cauchy principal value, the instantaneous amplitude

$$a(t) = \{x^2(t) + \mathrm{HT}^2[x(t)]\}^{\frac{1}{2}}, \tag{59}$$

the instantaneous phase

$$\phi(t) = \arctan\left\{\frac{\mathrm{HT}[x(t)]}{x(t)}\right\}. \tag{60}$$

The instantaneous frequency can be estimated as the local derivative of the instantaneous phase

$$\omega(t) = \frac{d\phi(t)}{dt}. \tag{61}$$

For this approach to generate a physically meaningful instantaneous frequency, the signal is required to satisfy some crucial necessary conditions: the signal is monocomponent, zero mean locally, and the waveform is symmetric about the local zero mean. Usually, the aforementioned adaptive mode decomposition algorithms in Section 3 can separate a signal into such mono-components [6]–[26].

In addition, Bedrosian and Nuttall theorems impose a furhter constraint: the Fourier spectra of the instantaneous amplitude $a(t)$ and the carrier signal $\exp[j\phi(t)]$ do not overlap. Otherwise, the AM variations will contaminate the FM part, and the instantaneous frequency is subject to the influence by the AM variations, resulting in occasional negative frequency values [27], [28].

2) EMPIRICAL AM-FM DECOMPOSITION

The mono-components obtained from adaptive mode decomposition algorithms do not satisfy the Bedrosian and Nuttall theorems automatically. To address this issue, Maragos *et al.* [29] proposed empirical AM-FM decomposition method.

The empirical AM-FM decomposition is based on iterative applications of cubic spline fitting to the local maxima of the signal absolute value, and separates the AM and FM parts of any IMF signal uniquely but empirically through a normalization scheme:

Step 1: Initialization: Given an IMF $c(t)$, set iteration index $i = 1$, residual signal $r_0(t) = c(t)$.

Step 2: For the residual signal $r_{i-1}(t)$, identify all the local maxima of its absolute value.

Step 3: Fit to the local maxima using a cubic spline, obtaining the empirical envelope $e_i(t)$.

Step 4: Normalize the residual signal $r_{i-1}(t)$ using the empirical envelope $e_i(t)$, obtaining the normalized residual $r_i(t) = \frac{r_{i-1}(t)}{e_i(t)}$. Through the absolute value data fitting, the normalized signal is guaranteed symmetric about the zero mean.

Step 5: Check the stop criterion, if $r_i(t) \leq 1$ for all t, terminate the normalization, and designate the empirical FM part $F(t) = r_i(t) = \cos[\phi(t)]$, which is a purely FM function with unity amplitude, and the AM part $a(t) = \frac{c(t)}{F(t)} = \prod_{k=1}^{i} e_k(t)$. Otherwise, let $i = i + 1$, go to step1 and repeat steps 2-5.

After empirical AM-FM decomposition, the IMF can be written as

$$c(t) = a(t)F(t) = a(t)\cos[\phi(t)]. \tag{62}$$

It separates any IMF empirically and uniquely into the corresponding instantaneous amplitude (AM) and the carrier signal (FM) parts. The obtained normalized carrier signal has unity amplitude, thus satisfying the Bedrosian and Nuttall theorem automatically, and enabling us to compute the direct quadrature.

The cubic spline fitted envelope serves as a better approach to the normalization operation. As a result, the empirical AM is smooth, and is devoid of the higher frequency fluctuation and overshoots. On the contrary, for the instantaneous amplitude obtained from the modulus of analytic signals, any nonlinear distortion in the raw signal waveform could cause even worse waveform deformation in the normalized signal.

The normalization process could cause some deformation of the raw signal waveform, but it is negligible, because: for a pure FM signal, its periodicity is mainly controled by the zero-crossings in addition to the extrema, and the zero-crossings are not altered in the normalization process.

3) DIRECT QUADRATURE

Direct quadrature avoids the Hilbert transform, which involves integral over time and is affected by neighboring data, thus enabling an exact estimation of instantaneous frequency [28]. The empirical FM signal derived from empirical AM-FM decomposition, is the carrier signal of the raw signal, and contains the instantaneous frequency informaiton. Suppose it is a cosine function $F(t) = \cos[\phi(t)]$, its quadrature can be derived as

$$\sin[\phi(t)] = \sqrt{1 - F^2(t)}. \tag{63}$$

There are two approaches to calculate the instantaneous phase based on the empirical FM signal: by taking the inverse cosine

$$\phi(t) = \arccos[F(t)], \tag{64}$$

or by taking the inverse tangent

$$\phi(t) = \arctan\left[\frac{\sqrt{1 - F^2(t)}}{F(t)}\right]. \tag{65}$$

Both approaches are based on differentiation only, and do not involve integral, thus preserving finely the instantaneous phase information of arbitrary form. As such, the resultant instantaneous phase is truly local, free from influences by any neighboring points. However, the former inverse cosine approach sometimes is unstable around the local extrema. The latter inverse tangent approach improves the computational stability. In addition, it uses all the four quadrants to uniquely calculate the phase angle, which is critical for proper phase unwrapping. To improve the stability at some possible

irregularities, the inverse tangent approach is modified with a median filter.

4) NORMALIZED HILBERT TRANSFORM

The empirical FM signal has identical unity amplitude, automatically satisfying the Bedrosian and Nuttall theorems. So the instantaneous phase can be calculated using the Hilbert transform based analytic signal method.

$$\phi(t) = \arctan\left\{\frac{\mathrm{HT}[F(t)]}{F(t)}\right\}. \tag{66}$$

The results from this approach are almost the same as the direct quadrature one, but they are different from those by the latter where the signal waveform has distortions. Such distortions are usually caused by changes in the instantaneous phase, violating the conditions imposed by the Bedrosian and Nuttall theorems. As such, only an approximate instantaneous frequency can be obtained via the normalized Hilbert transform (NHT) [28].

5) GENERALIZED ZERO-CROSSING

Zero-crossing approach estimates local frequency as half the inverse of interval between consecutive zero-crossings. It is the most fundamental one for local frequency estimation, but the result is constant over the period between zero-crossings, leading to a crude temporal resolution. To improve the performance, Huang *et al.* [28] took both the zero-crossings and the local extrema as the critical control points, and proposed the generalized zero-crossing (GZC) approach. This can improve the temporal resolution to a quarter wave period.

In the generalized scenario, we can obtain seven frequency values in three different classes. For the first class, the time interval between two consecutive critical control points of exactly the same type is considered as an entire wave period. For example, the interval between two consecutive up (or down) zero-crossings or two consecutive maxima (or minima) can be counted as one whole period. Such defined whole wave periods exactly cover a combination of all the four types of critical control point. As such, we may have four different period values at each time, when we consider a whole period with different critical control point as the starting point. We denote such whole periods as T_{4i}, where $i = 1, \cdots 4$.

For the second class, the interval between any two consecutive zero-crossings (from an up zero-crossing to next down one, or from a down zero-crossing to next up one), or any two consecutive extrema (from a maximum to next minimum, or from a minimum to next maximum), can be counted as a half period. Such defined half wave periods exactly cover two critical control points of the same kind, either the zero-crossings or local extrema. Therefore, we may have two different half period values at any time, when we take different critical control point as the starting point. We write such half periods as T_{2i}, where $i = 1, 2$.

For the third class, the interval between any two consecutive critical control points (from an extreme to next zero-crossing, or from a zero-crossing to next extreme) is considered as a quarter period. Such defined quarter wave periods exactly cover a combination of both extrema and zero-crossings yet only one from each type. Thus, we can have only one quarter period value at any instant, when we view different critical control point as the starting point. We express such quarter period as T_1.

At a given time, we calculate the mean frequency based on a weighted sum of the above three classes of period as

$$\begin{aligned}\bar{\omega} &= \frac{2\pi}{w_1 + 2w_2 + 4w_4}\left(\frac{w_1}{4T_1} + \sum_{i=1}^{2}\frac{w_2}{2T_{2i}} + \sum_{i=1}^{4}\frac{w_4}{T_{4i}}\right) \\ &= \frac{\pi}{6}\left(\frac{1}{T_1} + \sum_{i=1}^{2}\frac{1}{T_{2i}} + \sum_{i=1}^{4}\frac{1}{T_{4i}}\right),\end{aligned} \tag{67}$$

where the quarter period T_1 has the best time localization, thus being assigned a weight of $w_1 = 4$. The half period T_{2i} is given a weight of $w_2 = 2$ because of its less time localization. The whole period T_{4i} is set a weight of $w_4 = 1$ for its least time localization.

The GZC approach is derived from fundamental frequency definition based on wave period, yet does not involve any transform or differentiation. Thus it is direct and robust, and produces the most physically meaningful mean local frequency. In terms of algorithm, it is easy to implement, once the mono-components are available. However, its time localization is crude, being local down to a quarter period at most. It is also unable to represent waveform distortions, because it does not admit harmonics and intra-frequency modulations.

6) ENERGY SEPARATION

The energy separation (ES) approach is effective in estimating both the instantaneous frequency and the instantaneous amplitude of arbitrary time-varying modulated signals. It does not need to construct any basis functions, and is a completely data-driven algorithm adaptive to the local structure of a signal. It has attractive features such as high time–frequency resolution, adaptability to instantaneous feature, and low computational complexity [29], [30].

Teager energy operator is a nonlinear differential operator. It can estimate the energy required to generate a signal by means of nonlinear combination of the instantaneous signal values and its derivatives [29], [30]. For any signal $x(t)$, the Teager energy operator Ψ is defined as

$$\Psi[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \tag{68}$$

where $\dot{x}(t)$ and $\ddot{x}(t)$ are the first and the second derivatives of $x(t)$ with respect to time t, respectively. Actually, the output of energy operator tracks the energy required to generate the signal $x(t)$. Its counterpart for discrete time signals is defined as

$$\Psi[x(n)] = [x(n)]^2 - x(n-1)x(n+1). \tag{69}$$

Equation (69) shows that the Teager energy operator only needs three samples to calculate the signal source energy at

TABLE 1. Root mean squared error (RMSE) of instantaneous frequency estimation.

Method	HT	DQ	NHT	GZC	ES
RMSE	0.0036	0.0099	0.0020	0.0645	0.0180

any time and thus it has a good adaptability to the instantaneous changes in signals and an excellent ability to resolve transient events.

For signals of slowly time-varying or constant amplitude and frequency, the absolute value of instantaneous amplitude $a(t)$ and the instantaneous frequency $\omega(t)$ can be estimated as

$$|a(t)| = \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}}, \tag{70a}$$

$$\omega(t) = \sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}}. \tag{70b}$$

B. ILLUSTRATION

To illustrate the aforementioned instantaneous frequency estimation approaches, we generate an AM-FM signal

$$x(t) = [1 + 0.9\cos(2\pi f_{\mathrm{AM}} t + \pi/3)] \times \cos[2\pi f_{\mathrm{carrier}} t + \cos(2\pi f_{\mathrm{FM}} t)], \tag{71}$$

where the signal carrier frequency $f_{\mathrm{carrier}} = 1000$ Hz, the modulating frequency of AM and FM parts $f_{\mathrm{AM}} = f_{\mathrm{FM}} = 80$ Hz, and the sampling frequency is 80000 Hz. This AM-FM signal is representative, because both its instantaneous amplitude and instantaneous frequency vary with time.

Fig. 12 shows the instantaneous frequency estimation results of the synthetic AM-FM signal, using the HT, DQ, NHT, GZC and ES approaches respectively, and Table 1 lists their estimation error. All the approaches identify the profile of the instantaneous frequency curve. HT has a small root mean squared error, but it suffers from some spikes, as shown in Fig. 12 (b). GZC shows some steps due to its low time resolution, as presented in Fig. 12 (e), resulting in the largest error. Fig. 12 (f) displays the ES result, it has some ripples, and its error is larger than that of HT. The imperfection, for example spikes, steps and ripples, in approximation to the instantaneous frequency would mislead further analysis, such as time–frequency representation and frequency demodulation. DQ generates a better result. Although its error is larger than that of HT, it perfectly approximates the true instantaneous frequency curve, as shown in Fig. 12 (c). NHT produce the best result in terms of both approximation error and fitting consistence with the true instantaneous frequency curve, as presented in Fig. 12 (d).

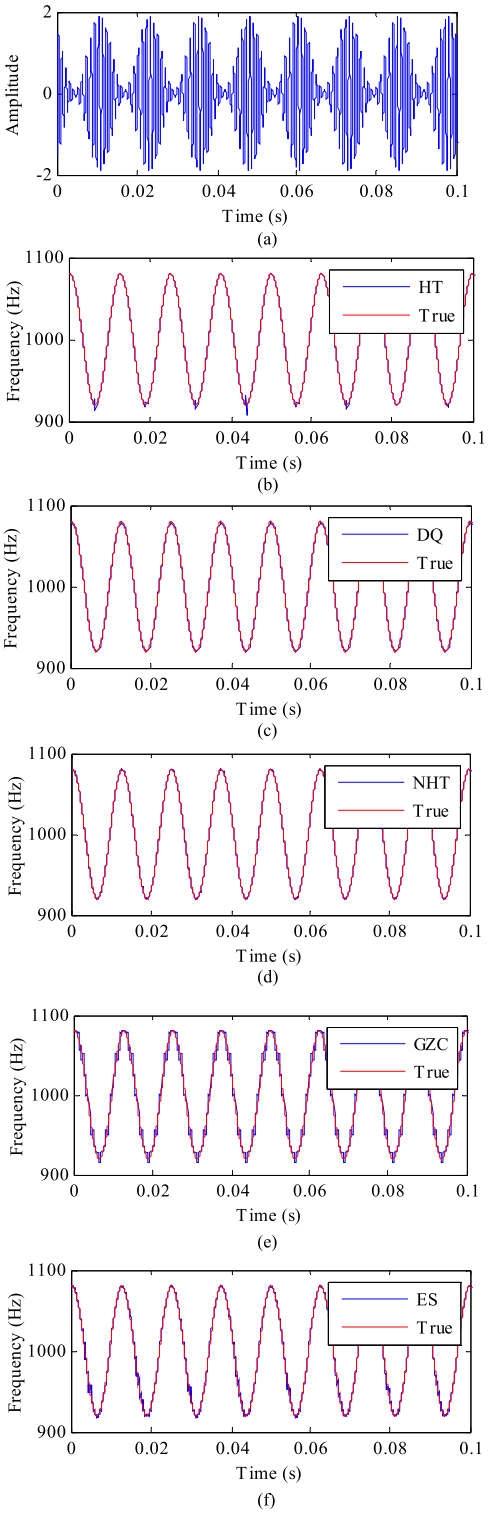

FIGURE 12. Instantaneous frequency estimation results. (a) Signal waveform; (b) HT; (c) DQ; (d) NHT; (e) GZC; (f) ES.

C. APPLICATION REVIEW

The instantaneous frequency estimation approaches have been applied in machinery fault diagnosis. Liang and Bozchalooi [85], [86] exploited the capability of the Teager energy operator in tracking transient changes in signals to detect impulses and thereby to diagnose bearing and gear faults. Ghazali *et al.* [87] compared the instantaneous frequency estimation approaches, including HT, NHT, DQ, energy separation and cepstrum approaches, and found that

the NHT, DQ and energy separation approaches are useful for leak detection of pipeline networks via pressure transient analysis. Wu *et al.* [88] proposed an instantaneous dimensionless frequency normalization method to characterize gear fault under variable running speed. They calculated instantaneous frequency based on NHT, GZC and DQ approaches, and then normalized the instantaneous frequency by the shaft rotating frequency to remove the effect of variable rotating speed. Furthermore, they extracted gear fault feature in joint time-dimensionless frequency distribution and marginal dimensionless frequency spectrum. Later, Wu *et al.* [89] extended the idea to rolling bearing fault diagnosis under variable speeds. Chen and Lin [90] calculated the instantaneous frequency and mean local frequency of IMFs obtained from EEMD based on DQ and GZC approaches respectively, and used the deviation of the instantaneous frequency from the mean local frequency to quantify the nonlinearity and nonstationarity of vehicle-track coupling systems for vehicle health monitoring. They showed the capability by case studies of wheels out-of-roundness and yaw damper failure.

D. REMARKS

All the above instantaneous frequency estimation approaches are based on mono-components. To satisfy the mono-component requirement, a multi-component signal should be decomposed into its constituent mono-components firstly. Usually, GZC provides the most stable local mean frequency, but cannot accurately track the instantaneous frequency of intrawave FM process due to the lower time resolution. The energy separation algorithm is also subject to a constraint that the instantaneous amplitude and instantaneous frequency of signals do not vary too fast or too greatly with time compared to the carrier frequency. Hilbert transform based analytic signal approach is widely used, but it requires the mono-component meet the Bedrosian and Nuttall theorems. The empirical AM-FM helps satisfy this condition, and make so derived NHT and DQ perform better. NHT is more stable than DQ, but DQ is more accurate than NHT.

V. APPLICATION EXAMPLES

In this section, we aim to demonstrate the performance of adaptive mode decomposition with some representative application examples using some typical methods. Considering the complexity issue existing with machinery dynamic signals due to the multi-component nature, strong nonstationarity and time-varying modulation, we have focused on adaptive mode decomposition algorithms like EEMD, ITD and VMD, as well as instantaneous frequency estimation approaches such as Hilbert transform and energy separation. We concentrate on both rotors, gears and rolling bearings as representative research targets, because their signals feature morphological diversity, such as harmonics, modulations and impulses. For each case, we select one appropriate method to demonstrate its performance in real applications.

A. ROTOR VIBRATION SIGNAL ANALYSIS VIA ENHANCED VMD BY ITERATIVE GENERALIZED DEMODULATION

Effective extracting rotating frequency harmonic components, and tracking their frequency and amplitude changes, are a key to success in rotating machinery condition monitoring and fault diagnosis, since these frequency components contain key information about a rotating machinery health. Under time variant conditions, particularly variable speed conditions, rotating frequency and its harmonic components exhibit modulation features and even overlap in frequency domain, because they change over time following the profile of variable speed yet at different changing rates. VMD can effectively separate modulation mono-components from a signal, and provides a potential approach for nonstationry rotating machinery vibration signal analysis. However, it is limited to signals without spectral overlaps only, because it is based on narrow-band properties of the AM-FM IMF definition and thereby requires all components separable in frequency domain. To address such a limitation issue with VMD, and generalize VMD to more general signals with spectral overlaps, we propose to improve VMD via iterative generalized demodulation [91]. Step 1: Through generalized demodulation, transform a target component into a component of an almost constant frequency without spectral overlaps with and thus separable from others. Step 2: Separate the target constant frequency component via VMD, and select it according to its center frequency. Step 3: By applying inverse generalized demodulation on the target constant frequency component, recover the original target component. Through iterative application of steps 1-3, each time with a demodulation phase function specially designed for a specific target component, all the constituent mono-component can be separated. Given mono-components, their respective instantaneous frequency can be calculated based on empirical AM-FM decomposition and NHT, and the time–frequency distribution of original signal can be constructed accordingly. Through time–frequency analysis, the frequency contents of a rotating machinery vibration signal and their time variability can be revealed effectively. We validated the proposed method by applying it to the rotor vibration signal of a real hydroturbine in a hydraulic power station during a shut-down transient process.

Fig. 13 shows the analysis results. In the time–frequency distribution derived from enhanced VMD via iterative generalized demodulation, Fig. 13 (a), all the prominent components are clearly identified, and the time variability of each component is well tracked, by virtue of good time–frequency readability. The rotating frequency and its harmonics up to the fourth order can be clearly seen throughout the shut-down process. Additionally, transient components appear around 4.5 Hz during the whole process. However, in the time-frequency distribution derived from original VMD, Figure 13 (b), only the rotating frequency is roughly discerned, against some surrounding noisy speckles. Its harmonics are severely twisted and even lost. For example, from 10 s to 20 s, the second to fourth harmonics are lost.

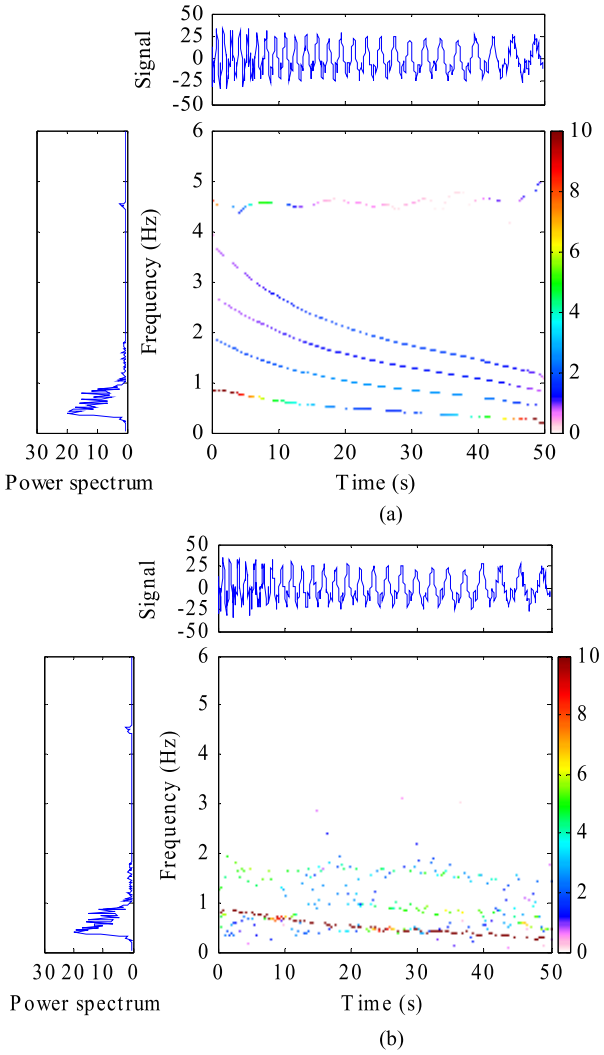

FIGURE 13. Time–frequency distribution of a hydroturbine vibration signal. (a) Enhanced VMD; (b) Original VMD.

From 20 s on, even though two rippling time-frequency ridges are coarsely revealed around the second and fourth harmonics, they deviate largely from the true ones and even cross each other. Additionally, transient components around 4.5 Hz are not discovered at all. This comparison demonstrates the outperformance of enhanced VMD over original VMD in processing nonstationary signals with spectral overlaps between constituent components.

B. PLANETARY GEAR FAULT DIAGNOSIS BASED ON ITD

Planetary gearbox vibration signals feature complicated modulations, thus leading to intricate sideband structure and resulting in difficulty in fault characteristic frequency identification. Intrinsic time–scale decomposition has unique merits, such as high adaptability to changes in signals, low computational complexity, good capability to suppress mode mixing and to preserve temporal information of transients, and excellent suitability for mono-component decomposition of complicated multi-component signals. In order to address the issue with planetary gearbox fault diagnosis due to the multiple modulation sources, a joint amplitude and frequency demodulation analysis method is proposed, by exploiting the merits of intrinsic time–scale decomposition. The signal is firstly decomposed into a series of mono-component proper rotational components, and the instantaneous frequency of each mono-component is calculated via the single wave based method, see Part C in Section III. Then the one with its instantaneous frequency fluctuating around the gear meshing frequency or its harmonics is selected as the sensitive component. Next, Fourier transformation is applied to the instantaneous amplitude and instantaneous frequency of the sensitive component to obtain the amplitude and frequency demodulated spectra respectively. Finally, a planetary gearbox fault is diagnosed by matching the peaks in the amplitude and frequency demodulated spectra with the theoretical gear fault characteristic frequencies. We validated the proposed method by analyzing the lab experimental signals of a planetary gearbox [92]. The localized faults of sun, planet and ring gears are diagnosed, showing the effectiveness of the method. Take the planet gear fault case as an example. Fig. 14 displays the analysis result. In the envelope spectrum of selected sensitive PRC and the Fourier spectrum of corresponding instantaneous frequency, Fig. 14 (b) and (c), prominent peaks emerge at the planet gear fault characteristic frequency and its harmonics mf_p, as well as their sum and difference combination with the planet carrier rotating frequency $mf_p \pm f_c$. Moreover, some peaks exist at the planet carrier rotating frequency and its harmonics nf_c, as well as the sun gear rotating frequency $f_s^{(r)}$. This is because the planet gear fault will result in an uneven load distribution among planet gears and thereby will magnify the AM effect of planet carrier rotation on gear meshing vibrations. These features accord with the planet gear fault symptom in amplitude and frequency demodulated spectra, thus implying the planet gear fault.

C. ROLLING BEARING FAULT DIAGNOSIS USING EEMD AND TEAGER ENERGY OPERATOR

Periodic impulses in vibration signals and their repeating frequency are the key indicators for rolling bearing localized defect detection. Teager energy operator is effective in detecting and highlighting impulsive components, while it requires the signal to be mono-component. EEMD has capability to decompose an intricate signal into mono-components. By virtue of EEMD and Teager energy operator, a new method is proposed to extract the characteristic frequency of rolling bearing fault. The signal is firstly decomposed into mono-components by means of EEMD to satisfy the mono-component requirement by Teager energy operator. Then the IMF of interest is selected according to strong correlation with the original signal and higher kurtosis value. Next Teager energy operator is applied to the selected IMF to detect fault induced impulses, by virtue of its good ability to highlight and detect transients. Finally Fourier transform is applied to the obtained Teager energy series to identify the repeating

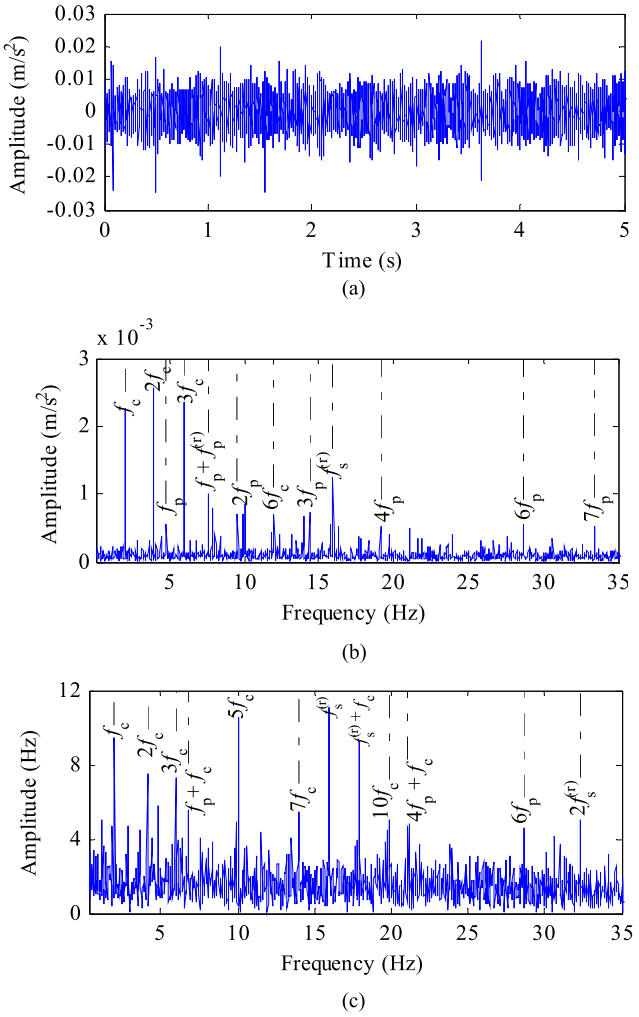

FIGURE 14. Planet gear fault signal. (a) Waveform; (b) Envelope spectrum; (c) Fourier spectrum of instantaneous frequency.

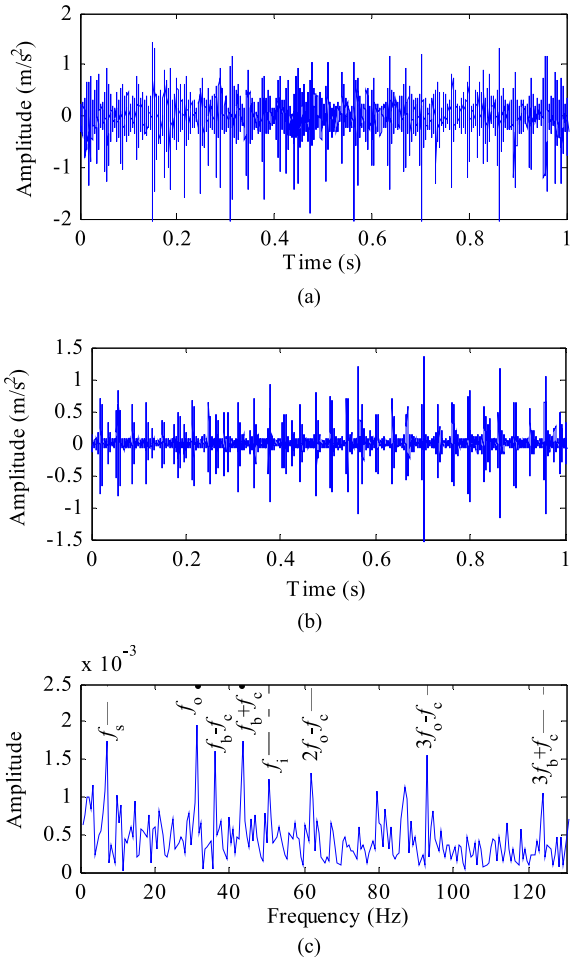

FIGURE 15. Rolling bearing compound fault signal. (a) Waveform; (b) Selected sensitive IMF waveform; (c) Teager energy spectrum of sensitive IMF.

frequency of fault induced periodic impulses, and thereby to diagnose rolling bearing faults. The proposed method has an excellent capability to finely detect transient impulses induced by localized defects, and thereby to resolve the periodicity of the impulse train. The resultant Teager energy spectrum is free of intricate sidebands, and can directly reveal the repeating frequency of periodic impulses characteristic of bearing faults, thus enabling easy interpretation of fault symptoms. We validated the proposed method by signal analyses of seeded fault experiments [93]. Fig. 15 shows the analysis results of a rolling bearing of compound faults with localized defect on its outer race, inner race and ball. In the Teager energy spectrum, Fig. 15 (c), prominent peaks are present, and the first most significant ones appear at the outer race characteristic frequency f_o, its harmonics minus the cage rotating frequency $2f_o - f_c$ and $3f_o - f_c$, the ball characteristic frequency and its harmonic plus/minus the cage rotating frequency $f_b \pm f_c$ and $3f_b + f_c$, and inner race characteristic frequency f_i, in addition to the shaft rotating frequency f_s. These frequencies relate to all the key component of rolling bearings, hence they indicate fault existence on the outer, inner races and ball, being consistent with the actual compound fault experimental settings.

VI. SUMMARY AND PROSPECTS

Adaptive mode decomposition is inspired by the idea behind EMD initiated in 1998. It is a remarkable conceptual progress from conventional orthogonal basis expansion to data-driven signal representation adaptive to arbitrary complicated signals. After nearly two decades of development, important achievements have been made on this method. Various adaptive mode decomposition algorithms and instantaneous frequency estimation approaches have been proposed, and their feasibility and capability in processing nonstationary complicated signals have been demonstrated in many reported studies from various fields.

This paper presents a systematic review of both adaptive mode decomposition algorithms and instantaneous frequency estimation approaches. These methods have their respective pros and cons, as summarized in Table 2 In real applications, one should select a proper method according to the specific characteristics of signals.

TABLE 2. Comparison of various adaptive mode decomposition methods.

Category	Method	Pros	Cons
Adaptive mode decomposition	EMD	Effective for most signals	Lack of rigorous mathematical formulation, subject to mode mixing, end effects, over/under shooting of cubic spline fitting, and higher sampling frequency demand
	LMD	Free from distortions by AM effects, positive instantaneous frequency	High computational complexity, prone to mode mixing
	ITD	Better suppresses end effect and mode mixing, avoids spikes in instantaneous frequency and negative frequency, low computational complexity	Low time resolution
	LCD	Higher computational efficiency, better mitigates mode mixing	Subject to end effects
	HVD	Mathematically rational, simple algorithm, low computational complexity, suited for quasi and almost periodic signals	Unsuited for impulsive and aperiodic signals
	EWT	Rigorous mathematical formulation, suited for constant frequency and almost periodic signals	Possible mode splitting, unsuited for highly nonstationary signals
	VMD	Rigorous mathematical formulation, suited for constant frequency and almost periodic signals	Possible mode splitting, unsuited for highly nonstationary signals, high computational complexity
	ALIF	Rigorous mathematical formulation, suited for constant frequency and almost periodic signals	Possible mode splitting, unsuited for highly nonstationary signals
	NMD	Rigorous mathematical formulation, best mono-component decomposition and time-frequency readability, suited for arbitrary complicated waveforms	High computational complexity
Instantaneous frequency estimation	HT	Suited for most signals	Subject to Nuttall and Bedrosian theorems, possible spikes in instantaneous frequency and negative frequency
	GZC	Meaningful mean local frequency, simple algorithm	Low time resolution
	ES	Simple algorithm, highly adaptive to transient changes	Subject to Nuttall and Bedrosian theorems, possible spikes in instantaneous frequency and negative frequency
	NHT	Free from Nuttall and Bedrosian theorems constraint, stable	Based on empirical AM-FM decomposition
	DQ	Free from Nuttall and Bedrosian theorems constraint, accurate	Based on empirical AM-FM decomposition

At present, adaptive mode decomposition has become an important research topic, and is attracting more and more researchers' attention. However, for real world complicated signal analysis in the field of machinery fault diagnosis, there are still some important issues to be investigated in-depth.

Except NMD, all the other adaptive mode decomposition methods are unable to effectively separate closely spaced frequency components. For signals composed of highly time-varying instantaneous frequencies and with spectral overlaps, particularly with instantaneous frequency crossings, nearly all the methods are unable to separate the constituent mono-components. Therefore, these methods are mostly effective for signal analysis under constant running conditions. They are also applicable to some certain nonstationary cases where running conditions change monotonically and slowly. How to generalize these methods to more general cases under arbitrary nonstationary running conditions is a key research topic, since machinery often works under variable speeds and the dominant rotating frequency and harmonics have spectral overlaps or even frequency crossings.

In mechanical engineering, signals are often contaminated by noise. This may lead to mode mixing in adaptive mode decomposition. Noise assistance has been demonstrated an effective approach to address the mode mixing issue in EEMD and ELMD. It would be an interesting and useful topic to study the effect of all kinds of noise on adaptive mode decomposition algorithms, and further develop corresponding noise assisted ensemble versions, to address the mode mixing issue.

It would be helpful to develop multivariate algorithms of adaptive mode decomposition. Usually, the number of modes obtained from an adaptive mode decomposition algorithm varies among different signals. As such, it is difficult to select a specific mode for comparison study among different signals, because of the incomparability among modes even with the same mode label. If multiple channel signals are treated as a multivariate signal, then the same number of modes can be produced via multivariate adaptive mode decomposition. This is useful to conduct some specific analyses by combining the modes with the same mode label, for example, full spectrum and even holo spectrum. Hence, extension of existing adaptive mode decomposition algorithms to multivariate version is an important research direction.

Regarding the characteristics of specific signals, how to select proper and true constituent components for further analysis is another important topic. Adaptive mode

decomposition usually produces more than one mode. Among all the obtained modes, some are true and contain the main information of interest, while the others might be pseudo and misleading. There still lacks a universal criterion to figure out the true and key modes of interest. In terms of signal feature extraction for machinery fault diagnosis, correlation of modes with original signal helps identify the true constituent ones, but cannot guarantee the selected one contain fault information. The instantaneous frequency of each mode is useful to assist mode selection, because fault signature is usually carried or manifested by specific frequencies. For example, gear damage information is often carried by the gear meshing frequency and its harmonics, rolling bearing defect feature is usually conveyed by the resonance frequencies, and rotor fault characteristics are commonly manifested by the rotating frequency and its harmonics. Therefore, it is suggested to include the instantaneous frequency as an additional index to select key and sensitive modes for further analysis.

Adaptive mode decomposition based time–frequency analysis is a better approach to nonstationarity analysis, transient detection, and time variability examination. For complicated multi-component signals, a quality time–frequency representation relies on fine mono-component decomposition and accurate instantaneous frequency estimation. Most of the current publications focus on early reported methods, such as the EMD and LMD decomposition algorithms and the HT estimation approach. Other methods have not been extensively studied. Therefore, it is necessary to investigate in-depth and exploit well the capabilities of both adaptive mode decomposition algorithms and instantaneous frequency estimation approaches, and construct quality time-frequency representations for complicated signal analysis in machinery fault diagnosis field.

Complicated signal analysis is a common yet key issue for machinery fault diagnosis. Appropriate signal analysis is important to identify the constituent components of signals and extract their features. Adaptive mode decomposition is highly adaptive and flexible in describing arbitrary signals, thus provides an effective approach to complicated signal analysis in machinery fault diagnosis.

ACKNOWLEDGMENT

The authors are grateful to N.E. Huang, Z. Wu, P. Flandrin, S. Smith, J. Cheng, M. Feldman, J. Gilles, K. Dragomiretskiy, D. Zosso, D. Iatsenko, and A. Cicone for sharing their adaptive mode decomposition codes. Comments and suggestions from all reviewers are appreciated.

REFERENCES

[1] Z. K. Peng and F. L. Chu, "Application of the wavelet transform in machine condition monitoring and fault diagnostics: A review with bibliography," *Mech. Syst. Signal Process.*, vol. 18, no. 2, pp. 199–221, Mar. 2004.

[2] R. Yan, R. X. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: A review with applications," *Signal Process.*, vol. 96, pp. 1–15, Mar. 2014.

[3] J. Chen *et al.*, "Wavelet transform based on inner product in fault diagnosis of rotating machinery: A review," *Mech. Syst. Signal Process.*, vols. 70–71, pp. 1–35, Mar. 2016.

[4] Z. Feng, M. Liang, and F. Chu, "Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples," *Mech. Syst. Signal Process.*, vol. 38, pp. 165–205, Jul. 2013.

[5] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mech. Syst. Signal Process.*, vol. 35, pp. 108–126, Feb. 2013.

[6] N. E. Huang *et al.*, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.

[7] N. E. Huang, Z. Shen, and S. R. Long, "A new view of nonlinear water waves: The Hilbert spectrum," *Annu. Rev. Fluid Mech.*, vol. 31, pp. 417–457, Jan. 1999.

[8] N. E. Huang *et al.*, "A confidence limit for the empirical mode decomposition and Hilbert spectral analysis," *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, vol. 459, pp. 2317–2345, Jul. 2003.

[9] G. Rilling, P. Flandrin, and P. Goncalves, "On empirical mode decomposition and its algorithms," in *Proc. IEEE-EURASIP Workshop Nonlinear Signal Image Process.*, Jun. 2003, pp. 8–11.

[10] Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, vol. 460, no. 2046, pp. 1597–1611, 2004.

[11] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, vol. 1, no. 1, pp. 1–41, 2008.

[12] J.-R. Yeh, J.-S. Shieh, and N. E. Huang, "Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method," *Adv. Adapt. Data Anal.*, vol. 2, no. 2, pp. 135–156, Apr. 2010.

[13] J. S. Smith, "The local mean decomposition and its application to EEG perception data," *J. Roy. Soc. Interface*, vol. 2, no. 5, pp. 443–454, Jul. 2005.

[14] M. G. Frei and I. Osorio, "Intrinsic time-scale decomposition: Time–frequency–energy analysis and real-time filtering of non-stationary signals," *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, vol. 463, no. 2078, pp. 321–342, Aug. 2006.

[15] J. Zheng, J. Cheng, and Y. Yang, "A rolling bearing fault diagnosis approach based on LCD and fuzzy entropy," *Mech. Mach. Theory*, vol. 70, pp. 441–453, Dec. 2013.

[16] M. Feldman, "Time-varying vibration decomposition and analysis based on the Hilbert transform," *J. Sound Vib.*, vol. 295, no. 3, pp. 518–530, Aug. 2006.

[17] M. Feldman, "Theoretical analysis and comparison of the Hilbert transform decomposition methods," *Mech. Syst. Signal Process.*, vol. 22, no. 3, pp. 509–519, Apr. 2008.

[18] M. Feldman, "Hilbert transform in vibration analysis," *Mech. Syst. Signal Process.*, vol. 25, no. 3, pp. 735–802, Apr. 2011.

[19] S. Braun and M. Feldman, "Decomposition of non-stationary signals into varying time scales: Some aspects of the EMD and HVD methods," *Mech. Syst. Signal Process.*, vol. 25, no. 7, pp. 2608–2630, Oct. 2011.

[20] M. Feldman, "Hilbert transform methods for nonparametric identification of nonlinear time varying vibration systems," *Mech. Syst. Signal Process.*, vol. 47, no. 1, pp. 66–77, Aug. 2014.

[21] M. Feldman and S. Braun, "Nonlinear vibrating system identification via Hilbert decomposition," *Mech. Syst. Signal Process.*, vol. 84, pp. 65–96, Feb. 2017.

[22] J. Gilles, "Empirical wavelet transform," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3999–4010, Aug. 2013.

[23] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 531–544, Feb. 2014.

[24] D. Iatsenko, P. V. McClintock, and A. Stefanovska, "Nonlinear mode decomposition: A noise-robust, adaptive decomposition method," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 92, no. 3, p. 032916, Sep. 2015.

[25] A. Cicone, J. Liu, and H. Zhou, "Adaptive local iterative filtering for signal decomposition and instantaneous frequency analysis," *Appl. Comput. Harmon. Anal.*, vol. 41, no. 2, pp. 384–411, Sep. 2016.

[26] L. Lin, Y. Wang, and H. Zhou, "Iterative filtering as an alternative algorithm for empirical mode decomposition," *Adv. Adapt. Data Anal.*, vol. 1, no. 4, pp. 543–560, Oct. 2009.

[27] A. H. Nuttall and E. Bedrosian, "On the quadrature approximation to the Hilbert transform of modulated signals," *Proc. IEEE*, vol. 54, no. 10, pp. 1458–1459, Oct. 1966.

[28] N. E. Huang, Z. Wu, S. R. Long, K. C. Arnold, X. Chen, and K. Blank, "On instantaneous frequency," *Adv. Adapt. Data Anal.*, vol. 1, no. 2, pp. 177–229, Apr. 2009.
[29] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. Signal Process.*, vol. 41, no. 4, pp. 1532–1550, Apr. 1993.
[30] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.
[31] D. Yu, J. Cheng, and Y. Yang, "Application of EMD method and Hilbert spectrum to the fault diagnosis of roller bearings," *Mech. Syst. Signal Process.*, vol. 19, no. 2, pp. 259–270, Mar. 2005.
[32] J. Cheng, D. Yu, J. Tang, and Y. Yang, "Application of frequency family separation method based upon EMD and local Hilbert energy spectrum method to gear fault diagnosis," *Mech. Mach. Theory*, vol. 43, no. 6, pp. 712–723, Jun. 2008.
[33] J. Cheng, D. Yu, J. Tang, and Y. Yang, "Local rub-impact fault diagnosis of the rotor systems based on EMD," *Mech. Mach. Theory*, vol. 44, no. 4, pp. 784–791, Apr. 2009.
[34] J. Zheng, J. Cheng, and Y. Yang, "Partly ensemble empirical mode decomposition: An improved noise-assisted method for eliminating mode mixing," *Signal Process.*, vol. 96, pp. 362–374, Mar. 2014.
[35] Z. K. Peng, P. W. Tse, and F. L. Chu, "A comparison study of improved Hilbert–Huang transform and wavelet transform: Application to fault diagnosis for rolling bearing," *Mech. Syst. Signal Process.*, vol. 19, no. 5, pp. 974–988, Sep. 2005.
[36] Z. K. Peng, P. W. Tse, and F. L. Chu, "An improved Hilbert–Huang transform and its application in vibration signal analysis," *J. Sound Vib.*, vol. 286, no. 1, pp. 187–205, Aug. 2005.
[37] S. J. Loutridis, "Damage detection in gear systems using empirical mode decomposition," *Eng. Struct.*, vol. 26, no. 12, pp. 1833–1841, Oct. 2004.
[38] B. Liu, S. Riemenschneider, and Y. Xu, "Gearbox fault diagnosis using empirical mode decomposition and Hilbert spectrum," *Mech. Syst. Signal Process.*, vol. 20, no. 3, pp. 718–734, Apr. 2006.
[39] R. Ricci and P. Pennacchi, "Diagnostics of gear faults based on EMD and automatic selection of intrinsic mode functions," *Mech. Syst. Signal Process.*, vol. 25, no. 3, pp. 821–838, Apr. 2011.
[40] G. Georgoulas, T. Loutas, C. D. Stylios, and V. Kostopoulos, "Bearing fault detection based on hybrid ensemble detector and empirical mode decomposition," *Mech. Syst. Signal Process.*, vol. 41, nos. 1–2, pp. 510–525, Dec. 2013.
[41] Z. Feng, H. Ma, and M. J. Zuo, "Amplitude and frequency demodulation analysis for fault diagnosis of planet bearings," *J. Sound Vib.*, vol. 382, pp. 395–412, Nov. 2016.
[42] Y. Lei, Z. He, and Y. Zi, "Application of the EEMD method to rotor fault diagnosis of rotating machinery," *Mech. Syst. Signal Process.*, vol. 23, no. 4, pp. 1327–1338, 2009.
[43] Y. Lei and M. J. Zuo, "Fault diagnosis of rotating machinery using an improved HHT based on EEMD and sensitive IMFs," *Meas. Sci. Technol.*, vol. 20, no. 12, p. 125701, Nov. 2009.
[44] Y. Lei, N. Li, J. Lin, and S. Wang, "Fault diagnosis of rotating machinery based on an adaptive ensemble empirical mode decomposition," *Sensors*, vol. 13, no. 12, pp. 16950–16964, Dec. 2013.
[45] J. Zhang, R. Yan, R. X. Gao, and Z. Feng, "Performance enhancement of ensemble empirical mode decomposition," *Mech. Syst. Signal Process.*, vol. 24, no. 7, pp. 2104–2123, Oct. 2010.
[46] M. Žvokelj, S. Zupan, and I. Prebil, "Multivariate and multiscale monitoring of large-size low-speed bearings using Ensemble Empirical Mode Decomposition method combined with Principal Component Analysis," *Mech. Syst. Signal Process.*, vol. 24, no. 4, pp. 1049–1067, May 2010.
[47] Z. Feng, M. Liang, Y. Zhang, and S. Hou, "Fault diagnosis for wind turbine planetary gearboxes via demodulation analysis based on ensemble empirical mode decomposition and energy separation," *Renew. Energ.*, vol. 47, pp. 112–126, Nov. 2012.
[48] D. Wang, W. Guo, and P. W. Tse, "An enhanced empirical mode decomposition method for blind component separation of a single-channel vibration signal mixture," *J. Vib. Control*, vol. 22, no. 11, pp. 2603–2618, Jun. 2016.
[49] A. Ayenu-Prah and N. Attoh-Okine, "A criterion for selecting relevant intrinsic mode functions in empirical mode decomposition," *Adv. Adapt. Data Anal.*, vol. 2, no. 1, pp. 1–24, Jan. 2010.
[50] H. Liang, Q.-H. Lin, and J. D. Z. Chen, "Application of the empirical mode decomposition to the analysis of esophageal manometric data in gastroesophageal reflux disease," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 10, pp. 1692–1701, Oct. 2005.
[51] Q. Gao, C. Duan, H. Fan, and Q. Meng, "Rotating machine fault diagnosis using empirical mode decomposition," *Mech. Syst. Signal Process.*, vol. 22, no. 5, pp. 1072–1081, Jul. 2008.
[52] M. Grasso, S. Chatterton, P. Pennacchi, and B. M. Colosimo, "A data-driven method to enhance vibration signal decomposition for rolling bearing fault analysis," *Mech. Syst. Signal Process.*, vol. 81, pp. 126–147, Dec. 2016.
[53] M. Grasso and B. M. Colosimo, "An automated approach to enhance multiscale signal monitoring of manufacturing processes," *J. Manuf. Sci. Eng.*, vol. 138, no. 5, p. 051003, Nov. 2015.
[54] Y. Yang, J. Cheng, and K. Zhang, "An ensemble local means decomposition method and its application to local rub-impact fault diagnosis of the rotor systems," *Measurement*, vol. 45, no. 3, pp. 561–570, Apr. 2012.
[55] J. Cheng, Y. Yang, and Y. Yang, "A rotating machinery fault diagnosis method based on local mean decomposition," *Digit. Signal Process.*, vol. 22, no. 2, pp. 356–366, Mar. 2012.
[56] Y. Wang, Z. He, and Y. Zi, "A demodulation method based on improved local mean decomposition and its application in rub-impact fault diagnosis," *Meas. Sci. Technol.*, vol. 20, no. 2, p. 025704, Jan. 2009.
[57] Y. Wang, Z. He, J. Xiang, and Y. Zi, "Application of local mean decomposition to the surveillance and diagnostics of low-speed helical gearbox," *Mech. Mach. Theory*, vol. 47, pp. 62–73, Jan. 2012.
[58] W. Y. Liu, W. H. Zhang, J. G. Han, and G. F. Wang, "A new wind turbine fault diagnosis method based on the local mean decomposition," *Renew. Energ.*, vol. 48, pp. 411–415, Dec. 2012.
[59] Z. Feng, M. J. Zuo, J. Qu, T. Tian, and Z. Liu, "Joint amplitude and frequency demodulation analysis based on local mean decomposition for fault diagnosis of planetary gearboxes," *Mech. Syst. Signal Process.*, vol. 40, no. 1, pp. 56–75, Oct. 2013.
[60] Z. Zheng, W. Jiang, Z. Wang, Y. Zhu, and K. Yang, "Gear fault diagnosis method based on local mean decomposition and generalized morphological fractal dimensions," *Mech. Mach. Theory*, vol. 91, pp. 151–167, Sep. 2015.
[61] Z. Liu, Z. He, W. Guo, and Z. Tang, "A hybrid fault diagnosis method based on second generation wavelet de-noising and local mean decomposition for rotating machinery," *ISA Trans.*, vol. 61, pp. 211–220, Mar. 2016.
[62] X. An, D. Jiang, J. Chen, and C. Liu, "Application of the intrinsic time-scale decomposition method to fault diagnosis of wind turbine bearing," *J. Vib. Control*, vol. 18, no. 2, pp. 240–245, Feb. 2012.
[63] X. An and D. Jiang, "Bearing fault diagnosis of wind turbine based on intrinsic time-scale decomposition frequency spectrum," *Proc. Inst. Electr. Eng. O, J. Risk Rel.*, vol. 228, no. 6, pp. 558–566, Jun. 2014.
[64] A. Hu, X. Yan, and L. Xiang, "A new wind turbine fault diagnosis method based on ensemble intrinsic time-scale decomposition and WPT-fractal dimension," *Renew. Energ.*, vol. 83, pp. 767–778, Nov. 2015.
[65] H. Liu, X. Wang, and C. Lu, "Rolling bearing fault diagnosis based on LCD–TEO and multifractal detrended fluctuation analysis," *Mech. Syst. Signal Process.*, vols. 60–61, pp. 273–288, Aug. 2015.
[66] Y. Qin, B. Tang, and Y. Mao, "Adaptive signal decomposition based on wavelet ridge and its application," *Signal Process.*, vol. 120, pp. 480–494, Mar. 2016.
[67] M. Kedadouche, M. Thomas, and A. Tahan, "A comparative study between Empirical Wavelet Transforms and Empirical Mode Decomposition Methods: Application to bearing defect diagnosis," *Mech. Syst. Signal Process.*, vol. 81, pp. 88–107, Dec. 2016.
[68] M. Kedadouche, Z. Liu, and V.-H. Vu, "A new approach based on OMA-empirical wavelet transforms for bearing fault diagnosis," *Measurement*, vol. 90, pp. 292–308, Aug. 2016.
[69] J. Pan, J. Chen, Y. Zi, Y. Li, and Z. He, "Mono-component feature extraction for mechanical fault diagnosis using modified empirical wavelet transform via data-driven adaptive Fourier spectrum segment," *Mech. Syst. Signal Process.*, vols. 72–73, pp. 160–183, May 2016.
[70] J. Chen, J. Pan, Z. Li, Y. Zi, and X. Chen, "Generator bearing fault diagnosis for wind turbine via empirical wavelet transform using measured vibration signals," *Renew. Energ.*, vol. 89, pp. 80–92, Apr. 2016.
[71] H. Cao, F. Fan, K. Zhou, and Z. He, "Wheel-bearing fault diagnosis of trains using empirical wavelet transform," *Measurement*, vol. 82, pp. 439–449, Mar. 2016.
[72] Y. Jiang, H. Zhu, and Z. Li, "A new compound faults detection method for rolling bearings based on empirical wavelet transform and chaotic oscillator," *Chaos, Solitons Fract.*, vol. 89, pp. 8–19, Aug. 2016.

[73] Y. Wang, R. Markert, J. Xiang, and W. Zheng, "Research on variational mode decomposition and its application in detecting rub-impact fault of the rotor system," *Mech. Syst. Signal Process.*, vols. 60–61, pp. 243–251, Aug. 2015.
[74] S. Zhang, Y. Wang, S. He, and Z. Jiang, "Bearing fault diagnosis based on variational mode decomposition and total variation denoising," *Meas. Sci. Technol.*, vol. 27, no. 7, p. 075101, Jul. 2016.
[75] Y. Wang and R. Markert, "Filter bank property of variational mode decomposition and its applications," *Signal Process.*, vol. 120, pp. 509–521, Mar. 2016.
[76] X. An and H. Zeng, "Pressure fluctuation signal analysis of a hydraulic turbine based on variational mode decomposition," *Proc. Inst. Mech. Eng., A, J. Power Energy*, vol. 229, no. 8, pp. 978–991, Sep. 2015.
[77] G. Tang, G. Luo, W. Zhang, C. Yang, and H. Wang, "Underdetermined blind source separation with variational mode decomposition for compound roller bearing fault signals," *Sensors*, vol. 16, no. 6, p. 897, Jun. 2016.
[78] Z. Lv, B. Tang, Y. Zhou, and C. Zhou, "A novel method for mechanical fault diagnosis based on variational mode decomposition and multikernel support vector machine," *Shock Vib.*, vol. 2016, Jan. 2016, Art. no. 3196465.
[79] C. Yi, Y. Lv, and Z. Dang, "A fault diagnosis scheme for rolling bearing based on particle swarm optimization in variational mode decomposition," *Shock Vib.*, vol. 2016, May 2016, Art. no. 9372691.
[80] H. Mahgoun, F. Chaari, and A. Felkaoui, "Detection of gear faults in variable rotating speed using variational mode decomposition (VMD)," *Mech. Ind.*, vol. 17, no. 2, p. 207, Mar. 2016.
[81] P. Clemson, G. Lancaster, and A. Stefanovska, "Reconstructing time-dependent dynamics," *Proc. IEEE*, vol. 104, no. 2, pp. 223–241, Feb. 2016.
[82] X. Zhang, J. Shao, W. An, T. Yang, and R. Malekian, "An improved time-frequency representation based on nonlinear mode decomposition and adaptive optimal kernel," *Elektron. Elektrotech.*, vol. 22, no. 4, pp. 52–57, 2016.
[83] X. An, H. Zeng, and C. Li, "Demodulation analysis based on adaptive local iterative filtering for bearing fault diagnosis," *Measurement*, vol. 94, pp. 554–560, Dec. 2016.
[84] X. An, H. Zeng, W. Yang, and X. An, "Fault diagnosis of a wind turbine rolling bearing using adaptive local iterative filtering and singular value decomposition," *Trans. Inst. Meas. Control*, p. 0142331216644041, Apr. 2016.
[85] M. Liang and I. S. Bozchalooi, "An energy operator approach to joint application of amplitude and frequency-demodulations for bearing fault detection," *Mech. Syst. Signal Process.*, vol. 24, no. 5, pp. 1473–1494, Jul. 2010.
[86] I. S. Bozchalooi and M. Liang, "Teager energy operator for multi-modulation extraction and its application for gearbox fault detection," *Smart Mater. Struct.*, vol. 19, no. 7, p. 075008, Jun. 2010.
[87] M. F. Ghazali, S. B. M. Beck, J. D. Shucksmith, J. B. Boxall, and W. J. Staszewski, "Comparative study of instantaneous frequency based methods for leak detection in pipeline networks," *Mech. Syst. Signal Process.*, vol. 29, pp. 187–200, May 2012.
[88] T. Y. Wu, J. C. Chen, and C. C. Wang, "Characterization of gear faults in variable rotating speed using Hilbert–Huang Transform and instantaneous dimensionless frequency normalization," *Mech. Syst. Signal Process.*, vol. 30, pp. 103–122, Jul. 2012.
[89] T. Y. Wu, C. H. Lai, and D. C. Liu, "Defect diagnostics of roller bearing using instantaneous frequency normalization under fluctuant rotating speed," *J. Mech. Sci. Technol.*, vol. 30, no. 3, pp. 1037–1048, Mar. 2016.
[90] S.-X. Chen and J.-H. Lin, "Nonlinearity and non-stationarity analysis of dynamic response of vehicle–track coupling system enhanced by Huang transform," *Measurement*, vol. 55, pp. 305–317, Sep. 2014.
[91] Z. Feng, F. Chu, and M. J. Zuo, "Time–frequency analysis of time-varying modulated signals based on improved energy separation by iterative generalized demodulation," *J. Sound Vib.*, vol. 330, no. 6, pp. 1225–1243, Mar. 2011.
[92] Z. Feng, X. Lin, and M. J. Zuo, "Joint amplitude and frequency demodulation analysis based on intrinsic time-scale decomposition for planetary gearbox fault diagnosis," *Mech. Syst. Signal Process.*, vols. 72–73, pp. 223–240, May 2016.
[93] Z. Feng, M. J. Zuo, R. Hao, F. Chu, and J. Lee, "Ensemble empirical mode decomposition-based Teager energy spectrum for bearing fault diagnosis," *J. Vib. Acoust.*, vol. 135, no. 3, p. 031013, Apr. 2013.

ZHIPENG FENG (S'89–M'89–SM'00) received the B.Sc. degree in automotive engineering from Jilin University in 1997, the M.Sc. degree in mechanical engineering from the Kunming University of Science and Technology in 2000, and the Ph.D. degree in power machinery engineering from the Dalian University of Technology, China, in 2003. From 2003 to 2005, he was a Postdoctoral Fellow at the Department of Precision Instruments and Mechanology, Tsinghua University, China. From 2006 to 2007, he was a Postdoctoral Research Fellow at the Department of Mechanical Engineering, University of Alberta, Edmonton, AB, Canada. He is currently a Professor at the School of Mechanical Engineering, University of Science and Technology Beijing, China. His research interests include machinery fault diagnosis, signal processing, artificial intelligence, and mechanical dynamics.

DONG ZHANG received the B.Sc. degree in mechanical engineering from the University of Science and Technology Beijing, China, in 2015, where he is currently working toward the M.Sc. degree in mechanical engineering. His current research interests include signal processing and machinery fault diagnosis.

MING J. ZUO (SM'00) received the B.Sc. degree in agricultural engineering from the Shandong Institute of Technology, China, in 1982 and the M.Sc. and Ph.D. degrees in industrial engineering from Iowa State University, Ames, IA, USA, in 1986 and 1989, respectively. He is currently a Professor at the Department of Mechanical Engineering, University of Alberta, Edmonton, AB, Canada. His research interests include system reliability analysis, maintenance modeling and optimization, signal processing, and fault diagnosis. He is a Fellow of the Institute of Industrial and Systems Engineering, a Fellow of the Engineering Institute of Canada, and a Founding Fellow of the International Society of Engineering Asset Management. He is an Associate Editor of the IEEE TRANSACTIONS ON RELIABILITY, the Department Editor of *IISE Transactions*, the Regional Editor for North and South American region for the *International Journal of Strategic Engineering Asset Management*, and an Editorial Board Member of *Reliability Engineering and System Safety*, the *Journal of Traffic and Transportation Engineering*, the *International Journal of Quality*, and the *International Journal of Performability Engineering*.