Hilbert–Huang Transform-Based Vibration Signal Analysis for Machine Health Monitoring

Ruqiang Yan, Student Member, IEEE, and Robert X. Gao, Senior Member, IEEE

Abstract—This paper presents a signal analysis technique for machine health monitoring based on the Hilbert-Huang Transform (HHT). The HHT represents a time-dependent series in a two-dimensional (2-D) time-frequency domain by extracting instantaneous frequency components within the signal through an Empirical Mode Decomposition (EMD) process. The analytical background of the HHT is introduced, based on a synthetic analytic signal, and its effectiveness is experimentally evaluated using vibration signals measured on a test bearing. The results demonstrate that HHT is suited for capturing transient events in dynamic systems such as the propagation of structural defects in a rolling bearing, thus providing a viable signal processing tool for machine health monitoring.

Index Terms—Empirical mode decomposition, Hilbert–Huang transform, intrinsic mode function, machine health monitoring, transient signal analysis.

I. INTRODUCTION

As a critical component in automated machine systems, effective machine health monitoring has attracted increasing attention from the research community over the past decade [1]–[4]. The goal of health monitoring is to identify potential failure causes at the early stage such that timely adjustment and maintenance actions can be taken to reduce severe machine damage and costly machine downtime. Of the various monitoring techniques commonly employed, vibration measurement remains an effective approach. This is due to the fact that structural failures in a machine system cause changes to the dynamic characteristics of the machine, which is reflected in its vibration signals. By means of appropriate signal decomposition and representation, features hidden in the vibration signals can be extracted, and assessment of the machine health status can be made.

Vibration signals encountered in rotary machine systems, such as machine tools or electric motors, can be broadly classified as being either stationary or nonstationary. Stationary signals are characterized by time-invariant statistic properties, e.g., the mean values and/or autocorrelation function [5]. Such signals can be adequately analyzed using well-known spectral techniques based on the Fourier Transform. An example is given in Fig. 1, where bearing misalignment caused by the inner and outer raceways falling out of the same plane resulted in periodic vibrations, which can be identified in its spectrum (see Fig. 1(b)). In contrast, nonstationary signals are “transient” in nature, with duration generally shorter than the observation interval. Such signals can be generated by the sudden breakage of a drilling bit, flaking of the raceway of a rolling bearing, or a growing crack inside a work piece. Effective detection of nonstationary signals is of great importance, as they are precursors of potential machine failures. However, their temporary
nature makes the assumption of signal stationarity as required by Fourier transform invalid, thus reducing its effectiveness.

In recent years, time-frequency analysis techniques, such as the Short-Time Fourier Transform (STFT) [6] and wavelet transform [7]–[9], have been investigated for their applicability in feature extraction from nonstationary, transient signals. Successful application of these techniques requires understanding of their respective limitations. For example, selection of a suitable window size is required when applying the STFT to match with the specific frequency content of the signal, which is generally not known a priori. When applying the wavelet transform, the type of the basic wavelet function employed directly affects the effectiveness in identifying transient elements hidden within the dynamic signal [10]. In comparison, the Hilbert–Huang transform (HHT) is based on the instantaneous frequencies resulting from the intrinsic mode functions of the signal being analyzed [11]–[13]; thus, it is not constrained by the uncertainty limitations with respect to the time and frequency resolutions to which other time-frequency techniques are subject. In recent years, HHT has been applied to identification of damage time instant and location in civil and mechanical structures [14]–[17]. Using HHT, the physical mass, damping coefficient, and stiffness matrices of a multiple degree of freedom (MDOF) linear system could be identified [18], [19]. In the area of biomedical engineering, studies on respiratory sinus arrhythmia from analyzing heartbeat time series [20] and heart rate variability [21] have further demonstrated the effectiveness of HHT in transient signal processing.

This paper investigates the utility of HHT as a tool for vibration signal analysis, with machine health monitoring as a target application area. After introducing the theoretical background, a synthetic signal is analytically formulated and used as a test signal for evaluating the performance of HHT. The simulation is subsequently evaluated through experimental studies performed on a bearing test bed. The results demonstrate the effectiveness of HHT for signal decomposition and feature extraction in machine health monitoring applications.

II. THEORETICAL BACKGROUND

The HHT represents the signal being analyzed in the time-frequency domain by combining the empirical mode decomposition (EMD) with the Hilbert transform [11]. In contrast to the Fourier spectral analysis by which a series of sine and cosine functions of constant amplitudes are used to represent each constituent frequency components in the signal, the HHT technique is based on the instantaneous frequency calculation that results from the Hilbert transform of the signal. Generally, the Hilbert transform $H[x(t)]$ for any signal $x(t)$ is defined as

$$H[x(t)] = y(t) = \frac{1}{\pi} \int \frac{x(\tau)}{t-\tau} \, d\tau$$

(1)

where $H[\bullet]$ denotes the Hilbert transform operation. Theoretically, any analytic signal $z(t)$ can be expressed by the sum of its real part $x(t)$ and imaginary part $y(t)$, with the latter being the Hilbert transform of the real part. This results in

$$z(t) = x(t) + jy(t)$$

(2)

Equation (2) can be rewritten in a polar coordinate system as

$$z(t) = a(t)e^{j\theta(t)}$$

(3)

where

$$\begin{align*}
\alpha(t) &= \sqrt{x(t)^2 + y(t)^2} \\
\theta(t) &= \tan^{-1}\left(\frac{y(t)}{x(t)}\right)
\end{align*}$$

(4)

represents the instantaneous amplitude and phase of the analytic signal, respectively. From the instantaneous phase $\theta(t)$, the instantaneous frequency $\omega(t)$ of the signal can be derived as

$$\omega(t) = \frac{d\theta(t)}{dt} = \frac{\dot{y}(t)x(t) - y(t)\dot{x}(t)}{x^2(t) + y^2(t)}.$$
Accordingly, the real part of the signal \( x(t) \) can be written in terms of the amplitude and instantaneous frequency as a time-dependent function

\[
x(t) = \Re(z(t)) = \Re(a(t)e^{j\int \omega(t)dt})
\]

where the symbol \( \Re(\cdot) \) denotes the real part of the analytic signal \( z(t) \).

To ensure that the instantaneous frequency obtained from the derivative operation in (5) is physically meaningful, the instantaneous phase \( \theta(t) \) must be a single-valued function over time, i.e., a mono-component function. Since the instantaneous phase \( \theta(t) \) is constructed from \( x(t) \) and its Hilbert transform, as shown in (4), it follows that \( x(t) \) must also be a mono-component function. The periodic vibration signal shown in Fig. 1 satisfies this requirement; thus, its instantaneous frequency can be calculated by using (3)–(5). The instantaneous frequency of 20 Hz is shown in the time-frequency representation (see Fig. 2) as a straight line across the data acquisition period, which is consistent with the time scale in Fig. 1.

To effectively construct frequency spectrum of a vibration signal that contains multiple-frequency components, the signal needs to be first decomposed into mono-component functions, by means of EMD [11]. A vibration signal measured from a ball bearing is shown in Fig. 3. Its spectrum reveals two components at 20 and 41 Hz, respectively, which are associated with the bearing misalignment and ball pass frequency in the outer raceway (caused by the periodic passing of the rolling balls over a fixed reference position).

Decomposition of such a signal is based on the following observations.

1) The signal has at least two extrema, i.e., one maximum and one minimum.

2) The characteristic time scale is clearly defined by the time lapse between successive alternations of local maxima and minima of the signal.

3) If the signal has no extrema but contains inflection points, then it can be differentiated one or more times to reveal the extrema.

The EMD technique decomposes the signal into a number of Intrinsic Mode Functions (IMFs), each of which a mono-component function. Then, the Hilbert transform is applied to calculate the instantaneous frequencies of the original signal.

The procedure for extracting the IMFs from a signal \( x(t) \) is illustrated in Fig. 4. After identifying all the local maxima and minima of the signal, the upper and lower envelopes are generated through curve fitting. Research has shown that many complex curve fitting functions have only resulted in marginal improvement while increasing the computational load significantly.
The mean values of the upper and lower envelopes of the signal \( m_{11}(t) \) are calculated as

\[
m_{11}(t) = \frac{(x_{\text{up}}(t) + x_{\text{low}}(t))}{2} \tag{7}
\]

where \( x_{\text{up}}(t) \) and \( x_{\text{low}}(t) \) are the upper and lower envelopes of the signal, respectively. Accordingly, the difference between the signal \( x(t) \) and the envelopes of the signal \( m_{11}(t) \), which is denoted as \( h_{11}(t) \), is given by

\[
h_{11}(t) = x(t) - m_{11}(t). \tag{8}
\]

Due to the approximation nature of the curve fitting method, \( h_{11}(t) \) has to be further processed (by treating \( h_{11}(t) \) as the signal itself and repeating the process continually) until it satisfies the following two conditions.

1) The number of extrema and the number of zero-crossings are either equal to each other or differ by at most one.

2) At any point, the mean value between the envelope defined by local maxima and the envelope defined by the local minima is zero.

Through the iteration process (for a total of \( k \) times), the difference between the signal and the mean envelope values, which is denoted as \( h_{1k}(t) \), is obtained as

\[
h_{1k}(t) = h_{1(k-1)}(t) - m_{1k}(t) \tag{9}
\]

where \( m_{1k}(t) \) is the mean envelope value after the \( k \)th iteration, and \( h_{1(k-1)}(t) \) is the difference between the signal and the mean envelope values at the \( (k-1) \)th iteration. The function \( h_{1k}(t) \) is then defined as the first IMF component and expressed as

\[
c_1(t) = h_{1k}(t). \tag{10}
\]

After separating \( c_1(t) \) from the original signal \( x(t) \), the residue is obtained as

\[
r_1(t) = x(t) - c_1(t). \tag{11}
\]
Subsequently, the residue $r_1(t)$ can be treated as the new signal, and the above-illustrated iteration process is repeated to extract the rest of the IMFs inherent to the signal $x(t)$ as

$$\begin{align*}
  r_1(t) - c_2(t) &= r_2(t) \\
  \vdots \\
  r_{n-1}(t) - c_n(t) &= r_n(t),
\end{align*}$$

(12)

The signal decomposition process is terminated when $r_n(t)$ becomes a monotonic function, from which no further IMFs can be extracted. By substituting (12) into (11), the signal $x(t)$ is decomposed into a number of intrinsic mode functions that are the constituent components of the signal. As a result, the signal $x(t)$ can be expressed as

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

(13)

where $c_i(t)$ represents the $i$th intrinsic mode function, and $r_n(t)$ is the residue of the signal decomposition. Equation (13) provides a complete description of the empirical mode decomposition process [11], which can be evaluated by checking the amplitude error between the reconstructed and the original signal. As an example, Fig. 5 illustrates both the IMFs and the residue of the multicomponent bearing vibration signal that is initially shown in Fig. 3. The reconstructed vibration signal is shown in Fig. 6, and the amplitude error between the original and reconstructed signals is calculated and illustrated in Fig. 6(d). Given that the maximum amplitude error is in the order of $10^{-16}$ (V), the EMD signal decomposition process can be considered to be complete.

Based on (1)–(5) and (13), (6) can be modified as

$$x(t) = \text{Re} \left( \sum_{i=1}^{n} a_i(t) e^{j \omega_i(t) t} \right)$$

(14)

in which $a_i(t) = \sqrt{c_i(t)^2 + H[c_i(t)]^2}$ and $\omega_i(t) = \frac{d}{dt} \left( \tan^{-1}(H[c_i(t)]/c_i(t)) \right)$. The term $r_n(t)$ in (13) is not included in (14) as it is a monotonic function and does not contribute to the frequency content of the signal.
Comparing (14) with the Fourier-based representation of a signal $x(t)$ given by

$$x(t) = \Re \left( \sum_{i=1}^{\infty} A_i e^{j \Omega_i t} \right)$$

(15)

where both $A_i$ and $\Omega_i$ are constant, it becomes evident that the EMD process enables flexible representation of a dynamic signal by revealing its time-dependent amplitude and the characteristic frequency components at various time instances. The signal is thus represented by a time-frequency distribution. The underlying HHT of the signal is mathematically defined as

$$\text{HHT}(t, \omega) = \sum_{i=1}^{n} \text{HHT}_i(t, \omega) = \sum_{i=1}^{n} a_i(t, \omega_i)$$

(16)

where $\text{HHT}_i(t, \omega)$ represents the time-frequency distribution obtained from the $i$th IMF of the signal. The symbol $\equiv$ denotes “by definition,” and $a_i(t, \omega_i)$ combines the amplitude $a_i(t)$ and instantaneous frequency $\omega_i(t)$ of the signal together.

In Fig. 7, the HHT result of the initial vibration signal (given in Fig. 3) is illustrated. The two major frequency components at 20 and 41 Hz were identified as constants throughout the observation period of 200 ms. Comparing to the Fourier spectrum, the time-frequency diagram has the advantage in revealing the time-dependency of the frequency components of interest. Such ability is particularly useful when analyzing transient signals whose constituent frequencies may change over time, e.g., when the health condition of a machine degrades due to defect growth and propagation.

In the above analysis, the test signal decomposed using the HHT process was experimentally measured from a realistic ball bearing. To generalize the concept of HHT-based vibration signal decomposition as a viable and effective signal processing tool for machine health monitoring, a quantifiable signal is needed to provide a computational reference base. For this purpose, a synthetic test signal was constructed, which consisted of defect-induced transient components with time-varying frequencies [10]. Amplitude variation of the transient components provides an indication of the bearing defect degradation. Given that defect-induced vibrations are generated every time when the rolling elements and the defect interact, the transient com-
Fig. 14. IMFs of vibration signals from two roller bearings. (Left) Healthy bearing. (Right) Defective bearing.

Components were formulated as a series of exponentially decaying oscillations \([10], [22]\]

\[
x(t) = \sum_{i=1}^{N} B_i \cdot e^{-C_i(t-t_i)^2} \cdot \sin(2\pi f_i (t - t_i) + \theta_i)
\]

with \(N\) being the number of transient elements, and \(B_i, C_i, t_i, \theta_i,\) and \(f_i\) corresponding to the amplitude, attenuation factor, time-delay, initial phase, and frequency of the \(i\)th transient component, respectively. The specific values of these parameters used for constructing the synthetic test signal are listed in Table I. In Fig. 8, the synthetic test signal is shown.

As seen in Fig. 8, a total of four groups of impulsive signal trains were created to simulate a transient vibration signal. Each of the pulse groups contains transient elements of two different center frequencies. For example, in the first two pulse groups, the center frequencies of the transient elements are equal to 1500 and 700 Hz, respectively. Similarly, the second two pulse groups were formulated using center frequencies at 1600 and 600 Hz. The four groups are separated from each other by a 12–ms time interval. Within each group, the two transient elements are time overlapped.

As shown in Fig. 8, the Fourier-based spectrum, although it reveals the four center frequencies contained in the signal, does not identify any frequency changes of the transient components. In comparison, the time-frequency decomposition of the signal based on HHT has revealed that between the time interval of 20–30 ms, which is the frequency composition of the signal changed from containing a 1500-Hz and a 700-Hz component to containing a 1600-Hz and a 600-Hz component. Such a result demonstrates the effectiveness of HHT in analyzing transient vibration components. The IMFs extracted during the EMD process are shown in Fig. 9, whereas in Fig. 10, the time-frequency decomposition of the various frequency components embedded in the synthetic test signal using HHT is illustrated.

To enable quantitative comparison between HHT and other time-frequency analysis techniques, the same signal was analyzed using the STFT and wavelet transform, as illustrated in Figs. 11 and 12 respectively. It can be seen that only when appropriate window size (i.e., 6.4 ms) and basic wavelet function (i.e., Morlet wavelet) were chosen could the transient components existing in the signal be successfully identified and differentiated. Such results reveal the constraints associated with these alternative time-frequency techniques, and comparatively indicate that HHT provides a viable approach to transient signal analysis.

### III. Experimental Evaluation

To experimentally evaluate the effectiveness of HHT for nonstationary vibration signal analysis, systematic tests were conducted on a custom-built bearing test bed (see Fig. 13). The test bed consisted of a DC motor (Lesson Cl6D34FT18), two supporting pillow blocks (SKF 209-112), a test bearing with supporting housing, a hydraulic cylinder (Miller 4Z644B), a hydraulic pump (Enerpac P392), and an optical encoder for
bearing speed measurement. The hydraulic cylinder applies variable load to the bearing in the radial direction. An accelerometer (bandwidth 1 Hz–30 kHz) measured vibrations from two test bearings (ball and roller bearing), which contained seeded defects on their raceways.

A. Defect Identification

The ability of the HHT for detecting bearing defect was first investigated on a roller bearing with a seeded defect (0.27 mm groove across its outer raceway). Shown in Fig. 14 is a comparison of extracted IMFs between vibration signals from a reference “healthy” bearing [without structural defect: Fig. 14 (left)] and a defective bearing [Fig. 14 (right)]. The corresponding results of the HHT analysis are illustrated in Fig. 15. For the defective bearing, transient vibrations caused by the rollers-defect interactions are clearly seen in the frequency range of 1–5 kHz and 8–12 kHz [see Fig. 15 (right). In addition, these transients have shown a repetitive pattern with a 6-ms interval, which corresponds to a 169-Hz repetitive BPFO frequency, resulting from the structural defect on the outer raceway. The healthy bearing, in comparison, showed no high-frequency components or repetitive signal patterns [see Fig. 15 (left), since no defect was present.

B. Defect Degradation Diagnosis

A run-to-failure experiment was conducted on a deep groove ball bearing of 52 mm outer diameter, in which a 0.27-mm-wide groove was cut across its outer raceway. Upon reaching approximately 2.7 million revolutions, the defect had propagated throughout the entire raceway and rendered the bearing practically nonfunctional. Vibration signals were taken during the course of the experiment, at an interval of every 7 min, to monitor the defect propagation process. Fig. 16 illustrates two data segments, covering the transitions of bearing test from phase I to II and from phase II to III.

The Fourier spectrum of the two data segments, shown in Fig. 17, illustrates averaged frequency information over the entire data sampling period, without providing specific information on defect propagation. In comparison, Fig. 18 illustrates the IMFs of the two data segments. The corresponding HHT identified frequency changes at the time instance of 45 ms [see Fig. 19 (left)] during the first data segment and at 55 ms during the second data segment [see Fig. 19 (right)]. Such frequency changes reflect degradation of the bearing health condition as the defect propagated through the bearing raceway. Physically, impacts generated by the rolling ball-defect interactions excite intrinsic modes of the bearing system, giving rise to a train of transient vibrations at the mode-related resonant frequencies. As the defect size increase, different intrinsic modes would be excited, resulting in the change of frequency components in the transient vibrations. In addition, these transients have shown a repetitive pattern with a 15-ms interval, which corresponds to a 67-Hz frequency component that is related to a bearing misalignment. Thus, the HHT has shown to provide an effective tool for bearing health diagnosis.

IV. CONCLUSION

As a time-frequency signal decomposition technique, the Hilbert-Huang transform provides an effective tool for analyzing transient vibration signal. Numerical and experimental studies on a custom-built bearing test bed have shown that
Fig. 17. Fourier spectrum of the two data segments. (Left) Segment 1. (Right) Segment 2.

Fig. 18. IMFs of the two data segments (Left) Segment 1. (Right) Segment 2.

Fig. 19. HHT of the two data segments. (Left) Segment 1. (Right) Segment 2.
the deterioration of a test bearing can be effectively detected through the time-dependent amplitudes and instantaneous frequencies resulting from the HHT. Such a technique can be further applied to the health monitoring of other types of dynamic systems, such as electrical drives.

The computational load of the HHT technique is determined by the EMD of the signal being analyzed, as it is an iteration process. To obtain a general idea on the computational efficiency of the HHT for vibration signal analysis, experiments were conducted on a laptop computer with a 2.0-GHz CPU and 1.0-GB memory. For the various signals investigated in the presented study, the computational time taken varied from 0.3 to 1.8 sec, which confirms the general applicability of the HHT technique for online machine health monitoring.

Research is being continued to systematically investigate the suitability and constraints of the HHT for nonstationary signal analysis, using vibration signals from different types of bearings. Furthermore, integration of the EMD process with enveloping spectrum analysis is also being investigated, with the goal of realizing an adaptive signal demodulation approach to account for varying machine conditions in real-world applications.

REFERENCES


Robert X. Gao (M’91–SM’00) received the B.S. and M.S. degrees in mechanical engineering from the University of Science and Technology of China, Hefei, China, in 1997 and 2002, respectively. He is currently working toward the Ph.D. degree at the Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst.

He is a Research Assistant at the Department of Mechanical and Industrial Engineering, University of Massachusetts. His research interests include nonlinear time-series analysis, multidomain signal processing, machine condition monitoring and health diagnosis, sensors, and sensor networks.

Dr. Gao received the National Science Foundation CAREER Award in 1996, is a faculty advisor and co-recipient of the inaugural Best Student Paper Award from the SPIE’s International Symposium on Smart Structures and Materials in 1997, and received the University of Massachusetts Outstanding Engineering Junior Faculty Award in 1999. He is a Co-Editor of the book Condition Monitoring and Control for Intelligent Manufacturing (Springer Verlag: 2006). He was the Guest Editor for a Special Section on Built-In-Test for the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, published in June 2005, and the Guest editor for a Special Section on Sensors for the Journal of Dynamic Systems, Measurement, and Control, published by the American Society of Mechanical Engineers (ASME) in June 2004. Currently, he serves as an Associate Editor for the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT and an Associate Editor for the ASME Journal of Dynamic Systems, Measurement, and Control. He is also on the Editorial Board of the International Journal of Manufacturing Research, He is a co-chair of the Technical Committee on Built-in-Test and Self-Test of the IEEE Instrumentation and Measurement Society.


Ruqiang Yan (S’04) received the B.S. and M.S. degrees in mechanical engineering from the University of Science and Technology of China, Hefei, China, in 1997 and 2002, respectively. He is currently working toward the Ph.D. degree at the Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst.

He is a Research Assistant at the Department of Mechanical and Industrial Engineering, University of Massachusetts. His research interests include nonlinear time-series analysis, multidomain signal processing, machine condition monitoring and health diagnosis, sensors, and sensor networks.

Robert X. Gao (M’91–SM’00) received the B.S. and M.S. degrees from the Technical University Berlin (TU Berlin), Berlin, Germany, in 1991, and 1985, respectively.

Presently, he is a Professor with the Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst. His research and teaching interests include physics-based sensing methodology, self-diagnostic and energy-efficient sensors and sensor networks, mechatronic systems design, medical instrumentation, and wavelet transforms for machine health monitoring, diagnosis, and prognosis.