Smart Sensors, Measurement and Instrumentation 26

Ruqiang Yan
Xuefeng Chen
Subhas Chandra Mukhopadhyay *Editors*

Structural Health Monitoring

An Advanced Signal Processing Perspective

Springer

Smart Sensors, Measurement and Instrumentation

Volume 26

Series editor

Subhas Chandra Mukhopadhyay
Department of Engineering, Faculty of Science and Engineering
Macquarie University
Sydney, NSW
Australia
e-mail: subhas.mukhopadhyay@mq.edu.au

More information about this series at http://www.springer.com/series/10617

Ruqiang Yan · Xuefeng Chen
Subhas Chandra Mukhopadhyay
Editors

Structural Health Monitoring

An Advanced Signal Processing Perspective

Springer

Editors
Ruqiang Yan
School of Instrument Science and Engineering
Southeast University
Nanjing
China

Xuefeng Chen
School of Mechanical Engineering
Xi'an Jiaotong University
Xi'an
China

Subhas Chandra Mukhopadhyay
Department of Engineering
Macquarie University
Sydney, NSW
Australia

ISSN 2194-8402 ISSN 2194-8410 (electronic)
Smart Sensors, Measurement and Instrumentation
ISBN 978-3-319-56125-7 ISBN 978-3-319-56126-4 (eBook)
DOI 10.1007/978-3-319-56126-4

Library of Congress Control Number: 2017936328

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The past decades have seen increasing attention from the research community worldwide on structural health monitoring (SHM). The efforts have promoted the continued advancement of sensing as well as signal processing technologies. In addition to commonly used time and frequency domain techniques, advanced signal processing techniques, such as wavelet transform and sparse representation, have been investigated as new tools for health monitoring of various mechanical or structural systems. However, many challenges and problems remain unsolved as of now or not fully addressed for SHM when signal processing techniques are applied to dealing with data measured from the system. For example, the complication of mechanical or structural systems results in complexity of the monitoring signals. Also, background noises weaken the effective condition signal and thus hinder the interpretation of the condition information. Furthermore, the specialization of each monitoring object leads to the predicament that a single signal processing technique cannot be effective for any SHM demands, which is also the reason why there are many advanced signal processing methods to be researched by academy and industry.

The book aims at introducing some advanced signal processing techniques that can be used in the field of structural health monitoring. The book contains invited chapters from researchers, who are experts in applying signal processing technique to solve structural health monitoring problems. It starts with an introduction on basic knowledge of structural health monitoring, followed by traditional frequency domain analysis, which is discussed for crack detection and rotor balance correction. Then some newly developed signal processing techniques, including wavelet transform, time-frequency analysis, compressive sensing and sparse representation, empirical mode decomposition, local mean decomposition and stochastic resonance, are introduced in theory with applications to various mechanical and structural systems. These advanced signal processing techniques are believed to be beneficial to structural health monitoring.

We would like to thank all the authors for their contribution and sharing of their knowledge. We do sincerely hope that the readers will find this book interesting and useful in their research on advanced signal processing for structural health monitoring.

Nanjing, China — Ruqiang Yan
Xi'an, China — Xuefeng Chen
Sydney, NSW, Australia — Subhas Chandra Mukhopadhyay

Contents

About the Editors

Dr. Ruqiang Yan (S'04-M'06-SM'11) received his Ph.D. degree from the University of Massachusetts Amherst in 2007, and his M.S. and B.S. degrees from the University of Science and Technology of China (USTC) in 2002 and 1997, respectively. He was a Guest Researcher at the National Institute of Standards and Technology (NIST) in 2006–2008. Dr. Yan joined the School of Instrument Science and Engineering at the Southeast University, China as a Professor in October 2009.

He is co-author of the book *Wavelets: Theory and Applications for Manufacturing* and has published over 100 refereed journal and conference papers. He was co-guest editor for special issues related to structural health monitoring in various journals. He received the *New Century Excellent Talents in University Award* from the Ministry of Education in China, in 2009.

His research interests include instrumentation design, nonlinear time-series analysis, multi-domain signal processing, and energy-efficient sensing and sensor networks for the condition monitoring and health diagnosis of large-scale, complex, dynamical systems.

Dr. Yan was an Instrumentation and Measurement Society (IMS) AdCom member (2014–2016). He is currently the Vice President for Technical & Standards Activities of the IMS. He is also co-chair of the Technical Committee (TC-7) on Signals and Systems in Measurement. He is an Associate Editor of the IEEE Transactions on Instrumentation and Measurement. He

received recognition of the transactions "Outstanding Reviewer of 2011", "2014 Outstanding Associate Editor", and "2015 Outstanding Associate Editor".

Dr. Xuefeng Chen (M'12) is Full Professor and Dean of School of Mechanical Engineering in Xi'an Jiaotong University, P.R. China, where he received his Ph.D. degree in Mechanical Engineering in 2004.

He works as the executive director of the Fault Diagnosis Branch in China Mechanical Engineering Society. Besides, he is also a member of ASME and IEEE, and the chair of IEEE the Xian and Chengdu Joint Section Instrumentation and Measurement Society Chapter.

He has authored over 100 SCI publications in areas of composite structure, aeroengine, wind power equipment, etc. He won National Excellent Doctoral Thesis Award in 2007, First Technological Invention Award of Ministry of Education in 2008, Second National Technological Invention Award in 2009, First Provincial Teaching Achievement Award in 2013, First Technological Invention Award of Ministry of Education in 2015, and he received National Science Fund for Distinguished Young Scholars in 2012 and was awarded as Science & Technology Award for Chinese Youth in 2013. Additionally, he hosted a National Key 973 Research Program of China as principal scientist in 2015.

Dr. Subhas Chandra Mukhopadhyay (M'97, SM'02, F'11) graduated from the Department of Electrical Engineering, Jadavpur University, Calcutta, India with a **Gold medal** and received the Master of Electrical Engineering degree from Indian Institute of Science, Bangalore, India. He has Ph.D. (Eng.) degree from Jadavpur University, India and Doctor of Engineering degree from Kanazawa University, Japan.

Currently he is working as Professor of Mechanical/Electronics Engineering and Discipline Leader of the Mechatronics Degree Programme of the Department of Engineering, Macquarie University, Sydney, Australia. He has over 26 years of teaching and research experiences.

His fields of interest include smart sensors and sensing technology, wireless sensor networks, Internet of Things, electromagnetics, control engineering, magnetic bearing, fault current limiter, electrical machines and numerical field calculation.

He has authored/co-authored over **400** papers in different international journals, conferences and book chapter. He has edited **15** conference proceedings. He has also edited **15** special issues of international journals as lead guest editor and **27** books with Springer-Verlag.

He was awarded numerous awards throughout his career and attracted over NZ $4.2 M on different research projects.

He has delivered **292** seminars including keynote, tutorial, invited and special seminars.

He is a **Fellow** of IEEE (USA), a **Fellow** of IET (UK) and a Fellow of IETE (India). He is a Topical Editor of IEEE Sensors Journal and an Associate Editor IEEE Transactions on Instrumentation. He has organized many international conferences either as general chair or technical programme chair. He is the Ex-**Chair** of the IEEE Instrumentation and Measurement Society New Zealand Chapter. He chairs the IEEE IMS Technical Committee 18 on Environmental Measurements. He is a Distinguished Lecturer of the IEEE Sensors Council for 2017–2019.

Advanced Signal Processing for Structural Health Monitoring

Ruqiang Yan, Xuefeng Chen and Subhas C. Mukhopadhyay

Abstract This chapter starts with an introduction on structural health monitoring (SHM) and emphasizes its importance for engineering systems. Then four different stages, i.e., operational evaluation, data acquisition, feature extraction and diagnosis and prognosis, involved in SHM are briefly discussed, followed by review of each signal processing technique used in SHM, which will be described in the book.

1 Introduction

The progress in science and technology has advanced development of engineering systems, such as aircraft, wind turbine and machine tools. Many of these existing systems are currently nearing the end of their original design life, and also new engineering systems are more and more complicated with high-precision. The process of implementing a damage identification strategy for aerospace, civil and mechanical engineering infrastructure is referred to as structural health monitoring (SHM) [1]. SHM is an engineering service to guarantee the performance of these systems so that they can detect the onset of damage at the earliest possible time and predict the remaining useful life of the engineering system in order to prevent failures and optimize resource allocation. Therefore the potential economic and life-safety implications of early damage detection in aerospace, civil and mechanical engineering systems have motivated a significant amount of research in SHM [2].

The guarantee of life-safety is always a strong motivation especially for the aircraft. On one hand, once an aircraft accident occurs, it is possible to result in fatal crash. As we can see in the Fig. 1, flight GE235 of TransAsia Airways, carrying 58

R. Yan (✉)
Southeast University, Nanjing, People's Republic of China
e-mail: ruqiang@seu.edu.cn

X. Chen
Xi'an Jiaotong University, Xi'an, People's Republic of China

S.C. Mukhopadhyay
Macquarie University, Sydney, Australia

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors, Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_1

Fig. 1 The air crash of TransAsia airways

people, crashed into a Taipei river on Feb. 4, killing at least 40 people. Thanks to the introduction of SHM, an improvement in maintenance and a decrease of structure-caused accidents would result in a global reduction of accidents of less than 10% [3]. On the other hand, systems based on SHM can assess their working conditions and prevent the fatal fault, such as engine health management (EHM) and health and usage monitoring systems (HUMS). EHM, as one of the main SHM applications, can be considered as a collection of capabilities from which building blocks can be drawn to create customized architectures that best meet individual user needs (see Fig. 2). In 2014, Allan J. Volponi, the chief scientist in Pratt and Whitney published a review paper on ASME to summarize and discuss the past, present and future trends of the gas turbine engine health management. Moreover, the reference stressed that the future state of EHM would need to integrate and balance its onboard and off-board capabilities to maximize cost benefit and operational reliability [4]. Likewise, HUMS is also one of typical SHM applications. HUMS is a combination of health monitoring and usage monitoring to provide accurate information regarding the condition of various flight critical components [5].

The economic motivation is also very strong. The economic impact with the effect of SHM is hard to evaluate. For example, Boeing attempted to transfer the company from the manufacturing enterprise to the service enterprise in order to achieve more economic benefits. It reported that about 40% or more of both metal and composite structure for a modern fighter aircraft can be saved on maintaining

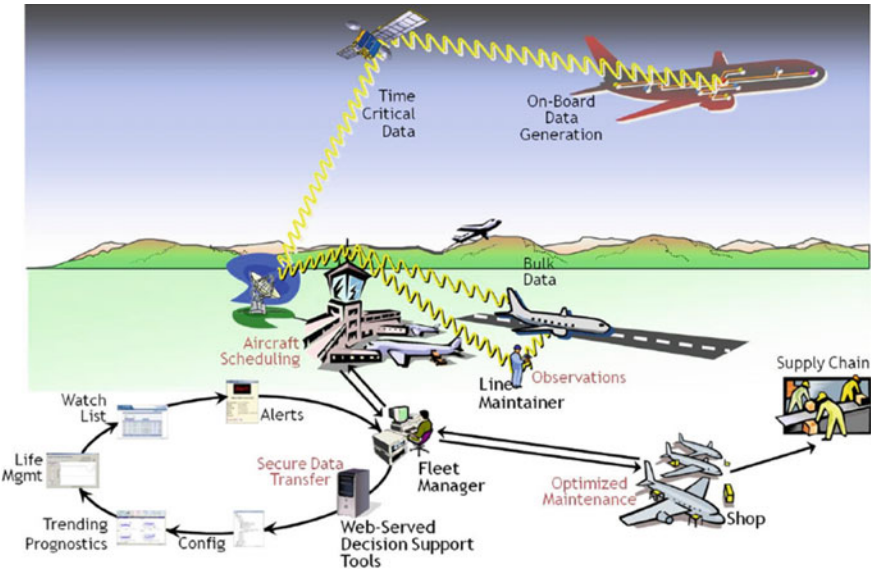


Fig. 2 EHM: the big picture [4]

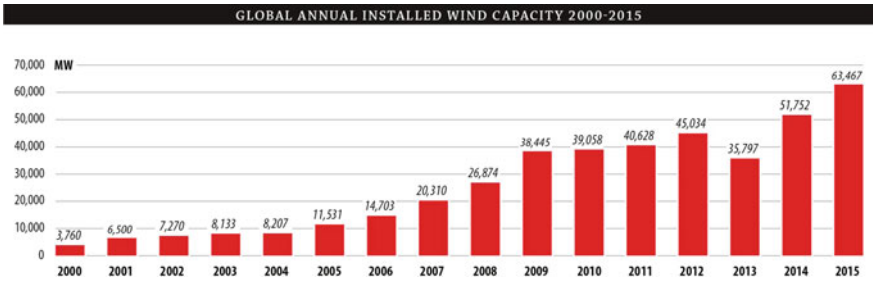


Fig. 3 Annual installed global capacity 2000–2015 [8]

time through the use of smart monitoring systems [6]. In terms of renewable energy, wind power is one of the fastest increasing renewable energy sources, and it is going to have remarkable share in the energy market [7]. With the installed capacity and scale of the wind turbines enlarging year by year (see Fig. 3; [8]), the economic benefits of wind power are seriously influenced by the maintenance costs and huge losses caused by frequent accidents of wind turbines. Thus, reducing maintenance costs has become an increasingly important issue, and SHM as an advanced monitoring technique can effectively help to solve this serious problem.

To sum up, the benefits of having SHM are as follows [9–11]:

(a) Enhancing structural safety: Based on the conditional information of SHM, enhancing the weak structure is feasible.

(b) Shortening unnecessary downtime: People can utilize the condition information to adjust the operation of the equipment.
(c) Avoiding catastrophic failures: Incipient fault detection can prevent the equipment from catastrophic failures and secondary defects.
(d) Reducing maintenance costs: The period maintenance can be replaced with condition-based maintenance.
(e) Supporting further development: Designer may yield detailed information on the dynamic behavior of the equipment over the periods of time that may help optimizing the design.

2 Structural Health Monitoring

Structural health monitoring (SHM) usually refers to the process of implementing a damage detection and characterization strategy for structural and mechanical systems. This book mainly focuses on the SHM of mechanical systems. Generally, SHM involves four different stages, namely: operational evaluation, data acquisition, feature extraction and diagnosis and prognosis (see Fig. 4).

2.1 Operational Evaluation

The process of operational evaluation is to answer the following four questions in order to implement the damage identification [12]:

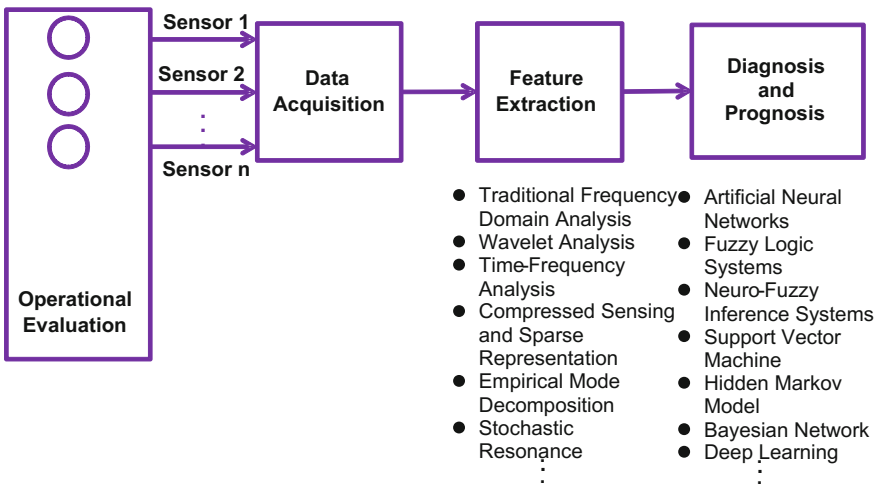


Fig. 4 Flow chart of SHM

- What is the life-safety or economic benefits for performing the SHM?
- What defines a damage for the system? And which cases are of the most concern?
- What are the operational and environmental conditions, under which the system to be monitored?
- What are the limitations on acquiring data under this special operational condition?

Operational evaluation mainly focuses on setting limitation on what should be monitored and how the monitoring can be finished. Namely, once the operational evaluation has been determined, the other component of the SHM would be carried out with increasing speed and high reliability.

2.2 *Data Acquisition*

Vibration measurement is an effective, reliable and non-intrusive technique to monitor machine conditions during startup, shutdown and normal operations [13]. Vibration sensors are one of the core components in a SHM system. It is no exaggeration to say that selection of sensors determines the accuracy of the signal and the reliability of SHM systems. Figure 5 shows the basic principle of data acquisition [14].

The vibration transducer detects the vibration parameter from the machine and converts these vibration signals to the electrical signals. The common used vibration transducers are:

- Displacement transducers (Vibrometer/Proximity Probes)
- Velocity transducers (Velometer)
- Accelerometers (PCB)
- Laser Doppler Vibrometers

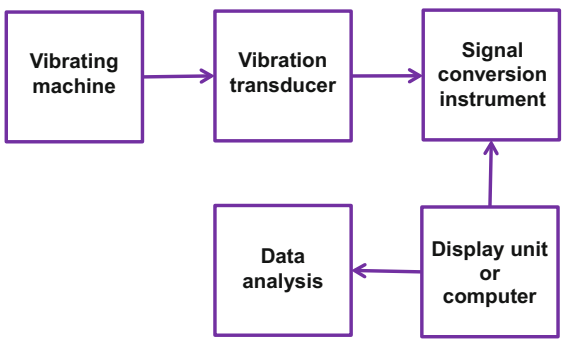


Fig. 5 Flow chart of data acquisition

Displacement transducers measure relative displacement which is suited to measure the low frequency vibration signal (below 10 Hz), whereas the most widely used to measure absolute motion are velocity (10–1000 Hz) and accelerometers (above 1000 Hz) which is suited to measure the high frequency vibration signal [15].

2.3 *Feature Extraction*

Extracting features through signal processing is the key component of any vibration based SHM, and is also the challenging aspect of the SHM. Because the acquired signals are always noisy and complex, the main aim is to remove noise and enhance the weak features of the system. Utilizing the features, the fault of the system can be identified. Moreover, location and severity of the damage will be determined for further diagnosis and prognosis.

Obviously, there are no universal physical variables and signal processing techniques that are appropriate for all the feature extraction problems [16]. Many signal processing techniques have been used to extract the features for SHM research. The most common ones are summarizes as follows [10, 17]:

- Frequency Domain Analysis: Fourier transform is used to estimate the strength of different frequency components contained in the signal. However, it cannot calculate the frequency at each instant of time, namely instant frequency.
- Wavelet Transform (WT) [18]: WT can achieve an optimal balance between frequency resolution and time resolution [19]. Moreover, it has multiple choices on basis function to match a specific fault symptom, which is beneficial to fault feature extraction [20, 21]. There are various types of wavelets used for SHM, such as discrete wavelet transform (DWT), continuous wavelet transform (CWT), wavelet packet transform (WPT) and second generation wavelet transform (SGWT) [22].
- Time-Frequency Analysis(TFA): TFA such as short-time Fourier transform (STFT), Chirplet transform, local polynomial Fourier transform and generalized demodulation approach [23–25] provides a powerful tool to effectively characterize the time-frequency (TF) pattern of nonstationary signals and gives an insight into the complex structure of a given signal consisting of several components.
- Compressed Sensing (CS) and Sparse Representation: Signal sparse representation is an advanced technique which utilizes a small number of atoms in the predefined dictionary to express the complex signal with a small error and it mainly consists of two basic procedures: dictionary construction and sparse approximation [26, 27]. In addition, compressed sensing brings a new inspiration to solve problems of big data compression, incomplete data processing and

rapid detection from small samples, which is regarded as a breakthrough of the Shannon sampling theorem [28, 29].

- Empirical Mode Decomposition (EMD): EMD is one of the most advanced signal processing techniques [30], which is proposed as an adaptive time-frequency signal processing method to analyze non-stationary and non-linear signals. It is based on the local characteristic time scales of a signal and could decompose the signal into a set of complete and almost orthogonal components called intrinsic mode function (IMF) [31].
- Stochastic Resonance (SR): SR is commonly described as an approach to increase the signal-to-noise ratio (SNR) at the output through the increase of the special noise level at input signal and is a nonlinear effect that is now widely used in weak signal detection under heavy noise circumstances [32, 33].

2.4 Diagnosis and Prognosis

In order to realize fault diagnosis, the quantification of the reliability and prognosis of remaining useful life, artificial intelligence (AI) is a necessary technique for SHM. SHM requires reliable models which are able to learn complex non-linear relationships between acquired signals and the system's state [34]. There are also various AI techniques to diagnose the machine state, evaluate its reliability and predict its remaining useful life. It is impossible to list all the AI techniques which can be used in the SHM system. According to the application, this portion lists some famous AI techniques for reference only.

- Artificial neural networks (ANNs): ANNs have powerful pattern classification and pattern recognition capabilities [35]. ANNs are appropriate to inspect the health conditions of machines due to its self-learning function, associated storage capabilities and abilities of fast searching an optimal solution.
- Fuzzy Logic Systems: Fuzzy logic is suitable for the representation of vague data and concepts on an intuitive basis, such as human linguistic description [36].
- Artificial Neuro-fuzzy Inference System (ANFIS): ANFIS integrates ANNs with fuzzy systems, and make use of their advantages to realize efficient diagnosis and prognosis [37].
- Support Vector Machine (SVM): SVM, based on statistical learning theory, is a famous machine learning method which is appropriate to small samples classification and regression [38]. Actually, SHM system confronts the problem of fewer samples, and SVM can handle this problem without loss of accuracy.
- Hidden Markov Models (HMMs): HMMs are attractive owing to their rich mathematical structure and their success in real world applications [39]. They can capture statistical properties of the underlying fault and result in reliable and efficient diagnosis and prognosis.

- Bayesian Networks (BN): BN is a model based on probability theory, and it describes how conditions are related through probabilities. BN and their extension for time-series modeling known as Dynamic Bayesian Network (DBN) have been shown by recent studies to be capable of providing a unified framework for SHM [40].
- Deep Learning (DL): DL refers to a class of machine learning techniques, where many layers of information processing stages in deep architectures are exploited for pattern classification and other tasks [41]. Based on DL, deep neural networks (DNNs) with deep architectures can be established to adaptively capture the representation information [42].

Among all the stages, the necessary step is to extract damage-related features through signal processing techniques to convert sensor data into damage information. Thus, signal processing techniques play a critical role in SHM.

3 Signal Processing in SHM

The goal of this book is to feature latest advances and directions in the advanced signal processing for SHM. The emphasis of the book will be on the utilization of advanced signal processing technique for helping to monitor the health status of some critical structures or equipment encountered in our daily life: wind turbine, gas turbine, machine tools, etc. Each chapter is organized as an introduction to an advanced signal processing technique for structural health monitoring, and gives a list of references, through which the readers can continue to research the state of the art signal processing techniques. A summary of key points in each chapter is given below.

The first part is about traditional frequency domain analysis. Chapter 2 shows a signal processing algorithm which can accurately estimate the natural frequencies of structures for early recognition and assessment of cracks. Chapter 3 improves the traditional holobalancing method by replacing the initial phase vector (IPV) with its forward precession component (IPV+). Thus, the impact of probe orientation on the balancing analysis and calculation is completely eliminated and the computational procedure is greatly simplified without sacrificing the balancing accuracy.

The second part is about the wavelet analysis. Chapter 4 verifies the essence on inner product operation of wavelet transform (WT) by simulation and field experiments. Moreover, the major developments on adaptive multiwavelet and super wavelet transform are introduced and discussed. Chapter 5 introduces a wavelet based spectral kurtosis and kurtogram that ensures automatic detection of impulsive transient vibrations occurring during machinery fault events.

The third part is about the time frequency analysis (TFA). Chapter 6 reports a new method called time-frequency manifold (TFM) which can effectively realize the signature enhancement and sparse representation of non-stationary signals for machinery fault diagnosis. Chapter 7 also proposes a new time-frequency analysis

method called matching demodulation transform (MDT) which can generate a time-frequency (TF) representation with satisfactory energy concentration, and thus extract the highly oscillatory frequency-modulation (FM) feature of rotating machine fault.

The forth part is about the compressive sensing and sparse representation. Chapter 8 brings a newly developed theory termed compressive sensing to the condition monitoring and fault diagnosis, and presents a novel method for rotary machinery fault detection from compressed vibration signals inspired by compressive sensing, which can largely reduce the data collection and detect faults of rotary machines from only a few signal samples. Chapter 9 presents an overview of the sparse representation theory, and utilizes the sparse representation theory to figure out the fault detection of rolling bearings, gearboxes and compound bearing faults.

The fifth part is about the empirical mode decomposition (EMD). Chapter 10 introduces the recent research and development of EMD in fault diagnosis of rotating machinery, and describes the basic concepts, fundamental theories and the applications about EMD methods and improved EMD methods. Chapter 11 discusses two dimensional form of EMD, namely Bivariate Empirical Mode Decomposition, and the powerful capacity of this innovative technique in the application of machine condition monitoring.

The final part is about two other state of the art methods including local mean decomposition and stochastic resonance. Chapter 12 proposes a time-frequency demodulation technique based on local mean decomposition which is utilized in extracting impulsive and modulated components of the rotor system and a gearbox. Chapter 13 focuses on the application of stochastic resonance (SR) in mechanical fault signal detection, and the ability of detecting weak signal is demonstrated through numerical simulations and experimental verification.

References

1. Farrar C.R., Worden K., "An introduction to structural health monitoring," Philosophical Transactions of the Royal Society A Mathematical Physical and Engineering Sciences, 2007, 365 (1851): 303–315.
2. Fugate M.L., Sohn H., Farrar C.R., "Vibration-based damage detection using statistical process control," Mechanical Systems and Signal Processing, 2001, 15 (4): 707–721.
3. Balageas D.L., Fritzen C., Guemes A., Structural Health Monitoring, John Wiley & Sons, Inc., 2006.
4. Volponi A.J., "Gas turbine engine health management: past, present, and future trends," Journal of Engineering for Gas Turbines and Power, 2014, 136 (5): 051201.
5. Samuel P.D., Pines D.J., "A review of vibration-based techniques for helicopter transmission diagnostics," Journal of Sound and Vibration, 2005, 282 (1): 475–508.
6. Bartelds G., "Aircraft structural health monitoring, prospects for smart solutions from a European viewpoint," Journal of Intelligent Material Systems and Structures, 1999, 9 (11): 906–910.

7. Farahani E.M., Hosseinzadeh N., Ektesabi M., "Comparison of fault-ride-through capability of dual and single-rotor wind turbines," Renewable Energy, 2012, 48(6): 473–481.
8. Council GWE, Annual Installed Global Capacity 2000–2015. http://www.gwec.net/global-figures/graphs/, 2015.
9. Caselitz P., Giebhardt J., Mevenkamp M., "On-line fault detection and prediction in wind energy converters," European Wind Energy Association Conference and Exhibition, 1994, pp. 623–627.
10. Goyal D., Pabla B.S., "The vibration monitoring methods and signal processing techniques for structural health monitoring: a review," Archives of Computational Methods in Engineering, 2015: 1–10.
11. Worden K., Farrar C.R., Manson G., Park, G., "The fundamental axioms of structural health monitoring," Proceedings of the Royal Society A, 2007, 463 (2082): 1639–1664.
12. Farrar C.R., Worden K., Structural Health Monitoring: A Machine Learning Perspective, John Wiley & Sons, Inc., 2012.
13. Girdhar P., Scheffer C., Practical Machinery Vibration Analysis and Predictive Maintenance, Elsevier, 2004.
14. Bishop R.E.D., Mechanical Vibrations, Allyn and Bacon, 1963.
15. Randall R.B., "Vibration-based condition monitoring: industrial, aerospace and automotive applications," John Wiley & Sons, Inc., 2010.
16. Wang L., Gao R.X., Condition Monitoring and Control for Intelligent Manufacturing, Springer London, 2006.
17. Amezquita-Sanchez J.P., Adeli H., "Signal processing techniques for vibration-based health monitoring of smart structures," Archives of Computational Methods in Engineering, 2016, 23 (1): 1–15.
18. Mallat S., A Wavelet Tour of Signal Processing: the Sparse Way, Academic Press, 2008.
19. Newland D.E., Wavelet Analysis of Vibration Signals, John Wiley & Sons, Inc., 2008.
20. Baccar D., Söffker D., "Wear detection by means of wavelet-based acoustic emission analysis," Mechanical Systems and Signal Processing, 2015, 60: 198–207.
21. Chen J., Li. Z, Pan J., Chen G., Zi Y., Yuan J., Chen B., He Z., "Wavelet transform based on inner product in fault diagnosis of rotating machinery: A review," Mechanical Systems and Signal Processing, 2016, 70: 1–35.
22. Yan R., Gao R.X., Chen X., "Wavelets for fault diagnosis of rotary machines: a review with applications," Signal Processing, 2014, 96: 1–15.
23. Qian S., Chen D., "Joint time-frequency analysis," IEEE Signal Processing Magazine, 1999, 16 (2): 52–67.
24. Sejdić E., Djurović I., Jiang J., "Time–frequency feature representation using energy concentration: An overview of recent advances," Digital Signal Processing, 2009, 19 (1): 153–183.
25. Wang S., Chen X., Cai G., Chen B., Li X., He Z., "Matching demodulation transform and synchrosqueezing in time-frequency analysis," IEEE Transactions on Signal Processing, 2014, 62 (1): 69–84.
26. Wright J., Yang A.Y., Ganesh A., Sastry S.S., Ma Y., "Robust face recognition via sparse representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31 (2): 210–227.
27. Elad M., Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing, Springer-Verlag, 2010.
28. Candès E.J., Wakin M.B., "An introduction to compressive sampling," IEEE Signal Processing Magazine, 2008, 25 (2): 21–30.
29. Donoho D.L., "Compressed sensing," IEEE Transactions on Information Theory, 2006, 52 (4): 1289–1306.
30. Huang N.E., Shen Z., Long S.R., Wu M.C., Shih H.H., Zheng Q., Yen N., Tung C.C., Liu H. H., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 1998: 903–995.

31. Lei Y., Lin J., He Z., Zuo M., "A review on empirical mode decomposition in fault diagnosis of rotating machinery," Mechanical Systems and Signal Processing, 2013, 35 (1): 108–126.
32. Hu N., Chen M., Wen X., "The application of stochastic resonance theory for early detecting rub-impact fault of rotor system," Mechanical Systems and Signal Processing, 2003, 17 (4): 883–895.
33. Benzi R., Sutera A., Vulpiani A., "The mechanism of stochastic resonance," Journal of Physics A: Mathematical and General, 1981, 14 (11): L453.
34. Abellan-Nebot J.V., Subirón F.R., "A review of machining monitoring systems based on artificial intelligence process models," International Journal of Advanced Manufacturing Technology, 2010, 47 (1–4): 237–257.
35. Zhang G., Patuwo B.E., Hu M.Y., "Forecasting with artificial neural networks: The state of the art," International Journal of Forecasting, 1998, 14 (1): 35–62.
36. Wu Y., Zhang B., Lu J., Du K.L., "Fuzzy logic and neuro-fuzzy systems: a systematic introduction," Journal of Materials Processing Technology, 2011, 129 (s 1–3): 148–151.
37. Wang WQ, Golnaraghi MF, Ismail F. Prognosis of machine health condition using neuro-fuzzy systems[J]. Mechanical Systems & Signal Processing, 2004, 18(4): 813–831.
38. Widodo A., Yang B.S., "Support vector machine in machine condition monitoring and fault diagnosis," Mechanical Systems and Signal Processing, 2007, 21 (6): 2560–2574.
39. Rabiner, L.R., "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, 1989, 77 (2): 257–286.
40. Iamsumang C., Mosleh A., Modarres M., "Computational algorithm for dynamic hybrid Bayesian network in on-line system health management applications," 2014 International Conference on Prognostics and Health Management, 2014, pp. 1–8.
41. Li D., "A tutorial survey of architectures, algorithms, and applications for deep learning," APSIPA Transactions on Signal and Information Processing, 2014, 3 e2:1–29.
42. Jia F., Lei Y., Lin J.,, Zhou X., Lu N., "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," Mechanical Systems and Signal Processing, 2015, 72–73: 303–315.

Signal Post-processing for Accurate Evaluation of the Natural Frequencies

G.R. Gillich and I.C. Mituletu

Abstract In this paper, a signal processing algorithm to accurately estimate the natural frequencies of structures for early recognition and assessment of cracks is proposed. In standard frequency estimation, the precision increases if the frequency resolution is improved. A finer resolution is achieved by increasing the analysis time interval. Nowadays, there are many other methods to improve the spectrum resolution, as the interpolation of spectral lines, zero-padding, zoom-FFT and so on. The proposed algorithm stepwise crops the acquired vibration signal and performs a spectral analysis. Superposing these spectra, an overlapped spectrum, with a dramatically increased resolution, results. This spectrum offers the possibility to identify very precisely the natural frequencies, even for damages in early stage. The algorithm was tested on generated and real-world signals and was proved to work well, even in the case of fast damped or short signals.

1 Introduction

Damage assessment, implying the location identification and severity estimation of cracks or other types of damage in structural members, using vibration data, has received considerable attention in last decades [1–3]. The basic idea is that modal parameters, such as frequencies or mode shapes, are functions of the physical parameters (e.g. stiffness, damping or mass). A common approach is to extract the modal parameters of the healthy structure as a baseline via modal identification methods [4], all subsequent test results being afterwards compared with this data [5]. Deviation of the modal parameters indicates the occurrence of damage. Thus, it is possible to find the location and magnitude of the structural alteration if the effect of the physical parameters changes upon the modal parameter is known [6].

G.R. Gillich (✉) · I.C. Mituletu
Department of Mechanical Engineering, University of Resita,
320085 Resita, Romania
e-mail: gr.gillich@uem.ro

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors,
Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_2

Numerous methods for damage assessment are presented in literature, differentiated by the level achieved in damage identification [7], the type of excitation [8], the number and the type of sensors involved in the structural monitoring process, the complexity of the structure's model and the algorithms applied to recognize and quantify the physical parameter changes from the modal parameter shifts. In engineering applications, the output-only damage detection methods are the most promising, because the excitations are usually difficult to measure. Therefore, it is desirable to involve detection methods based only on the measured responses, without making use of excitation information [9].

In order to be seriously considered for in situ implementation, damage detection methods should demonstrate that they can perform well under several limitations. For instance, it is important to achieve good results with a small number of measurement points, which should be selected a priori without knowledge of the damage location. Another important issue is the possibility to discriminate changes in the modal parameters due to damage from these resulting from variations in the environmental and/or test conditions. Also important is the repeatability of the tests; a reduced level of uncertainty in the measurements ensures the detection of damage in an early stage. Because it fulfils the above-enounced requirements, the natural frequency becomes the mostly utilized parameter for damage detection. Still, frequency changes present low sensitivity to damage, so that frequency-based methods require precise frequency evaluation.

2 Motivation

Damage detection methods based on the change in the natural frequencies are performed by acquiring and processing vibration signals. The time-domain signal is converted into the frequency domain by specific algorithms. For most structures the range between the harmonics is sufficiently large so that it is easy to discern between consecutive frequencies. However, problems in observing early damage occur because small damages cause limited changes in the natural frequencies. This happens because, in standard frequency evaluation, the frequencies are indicated at equidistantly distributed spectral lines, whose position is dependent on the signal length. Thus, for small structural parameter alterations, the frequency changes are not indicated in the spectrum since the peak-amplitude moves to the inferior spectral line. It results in the requirement of involving certain advanced algorithms providing denser spectral lines and in this way increasing accuracy when frequencies are evaluated.

In this work, we introduce a simple frequency evaluation algorithm and test it against several simple methods presented in the literature. The application of this algorithm, proved as exact and robust, will help in improving the early damage recognition and precise location and severity estimation.

3 Standard Frequency Evaluation

Vibration signals are continuous functions of time, thus analogue, as well as most sensor outputs. The accelerometer presented in Fig. 1 produces an output signal that is proportional to the acceleration of the system at the measurement coordinate. This analogue time signal, depicted in Fig. 2a, is transmitted from transducer to an electronic analogue-to-digital (A/D) converter which transforms it into a discrete time series, resulting in a digital signal. Now, the digital code can be used by the processor [10].

The A/D converter records the level of the signal at a discrete set of times. Figure 2 illustrates the sampled acquisition of an analogue time signal, where the time τ elapsed between each sample is

$$\tau = 1/F_S \tag{1}$$

if F_S denotes the sampling frequency. In the case of the National Instruments acquisition module NI 9233, shown in Fig. 1, the sampling frequency is

$$F_S = 50{,}000/\mathrm{m}\,[\mathrm{Hz}] \tag{2}$$

with m an integer less than 25. In practice, a collection of points with information on the amplitude at discrete and regular intervals of times is achieved, as shown in Fig. 2b.

A/D conversion of a signal $x(t)$ means to ensure an amplitude value for each discrete time $k{\cdot}\tau$, with k = 0, 1, 2, …N−1. If the individual amplitudes are denoted $x[k] = x(k\tau)$, the signal $x(t)$ is represented by following sequence:

$$\{x\} = \{x[0], x[1], \ldots, x[k], \ldots, x[N-1]\} \tag{3}$$

The process of converting a continuous sinusoid of amplitude a and frequency f into a discrete signal follows:

$$x(t) = a\sin(2\pi ft + \phi) \rightarrow x[\tau] = x(k\tau) = a\sin(2\pi fk\tau + \phi) \tag{4}$$

Fig. 1 The analogue-to-digital converter NI 9233 and a piezoelectric PCB accelerometer

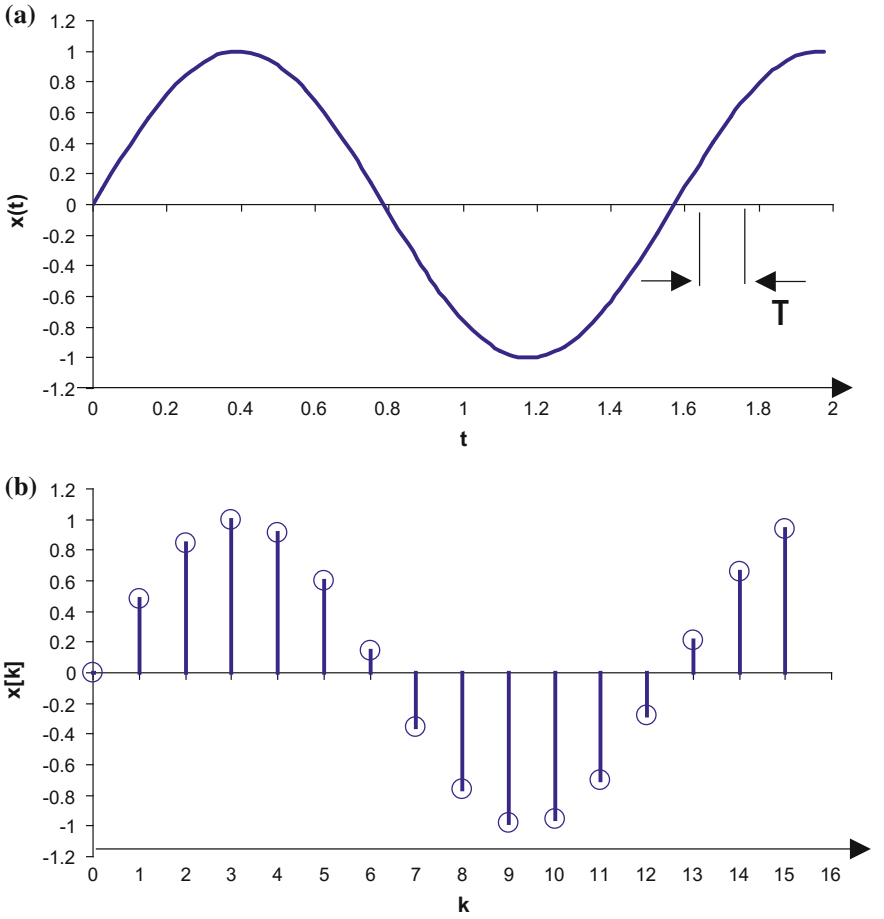


Fig. 2 The analogue and converted digital signal

The sequence $\{x\}$ is also called the sampled version of $x(t)$. It contains no explicit information about the sampling frequency (or sampling rate) F_S. However, because the number of samples N and the sampling time τ are known, the signal length (or acquisition time) is derived as

$$T_S = (N-1)\tau = \frac{(N-1)}{F_S}. \tag{5}$$

The conversion to discrete signals is in direct relation to the chosen sampling time τ respectively the sampling frequency F_S. A perfect signal reconstruction from the sampled values is possible, so Nyquist-Shannon sampling theorem, if the signal frequency does not exceed the half of the sampling frequency, i.e. $f < F_S/2$.

Any complex signal can be synthesized from a linear combination of harmonic functions. The inverse process, the analysis, permits decomposition of complex signals in harmonic components. One of the most utilized algorithms to represent a function in the frequency domain, aiming to highlight the harmonic components, is the Discrete Fourier Transform (DFT). It processes discrete signals of finite time, being ideal for processing information stored in computers.

For multi-frequency signals, the sequence in Eq. (3) can be expressed as a sum of complex sinusoids

$$x[k] = \sum_{j=0}^{N-1} a_j e^{i2\pi \frac{k}{N-1} j} \tag{6}$$

where j is the sinusoid index, and, by definition

$$a_j = \frac{1}{N} \sum_{j=0}^{N-1} x[k] e^{-i2\pi \frac{k}{N-1} j} \tag{7}$$

Let the continuous time function $x()$ be an approximation to the original continuous time function from which the $\{x\}$ sequence was obtained

$$x(k\tau) = \sum_{j=0}^{N-1} a_j e^{i2\pi \frac{k}{N-1} j} \quad \text{for } x \in R \tag{8}$$

or

$$x(t) = \sum_{j=0}^{N-1} a_j e^{i2\pi \frac{j}{N-1\tau} t} = \sum_{j=0}^{N-1} a_j e^{i2\pi f_j t} \tag{9}$$

where f_j is the frequency of the jth component, defined as

$$f_j = \frac{j}{(N-1)\tau} = \Delta f \cdot j \tag{10}$$

From Eq. (10) one can observe that the frequencies for which the analysis is performed are equidistantly distributed, the interval between two consecutive frequencies, called frequency resolution, being

$$\Delta f = \frac{1}{(N-1)\tau} = \frac{1}{T_S} \tag{11}$$

The DFT of the discrete sequence $\{x\}$ will indicate, at the jth spectral line, the amplitude

$$X[j] = Na_j = \sum_{k=0}^{N-1} x[k] e^{-i2\pi \frac{j}{N-1} k} \tag{12}$$

in the frequency domain. If frequencies f_j are considered in stand of spectral line numbers j, Eq. (12) becomes

$$X[f_j] = \sum_{k=0}^{N-1} x[k] e^{-i2\pi f_j t} \tag{13}$$

As a consequence, an input signal with N samples result in a DFT will N spectral lines. In Fig. 3, constituting the frequency spectrum, the individual values of $X[j]$ for $j = 0 \ldots N - 1$ spectral lines are indicated. One can observe that the number of samples in both the time and frequency representations is the same. Equation (13) shows that regardless of whether the input signal $x[k]$ is real or complex, $X[j]$ is always complex [11]. For real signals, such as those obtained from the output of a DAQ device, the DFT is symmetric, with the following property

$$|X[j]| = |X[N - j]| \tag{14}$$

Obviously, only half of the spectral lines of the DFT need to be computed or displayed, because the symmetry makes the information redundant.

Having a look onto Fig. 3 one can observe that it is sufficient to analyze the DFT only for $N/2$ samples, corresponding to a frequency

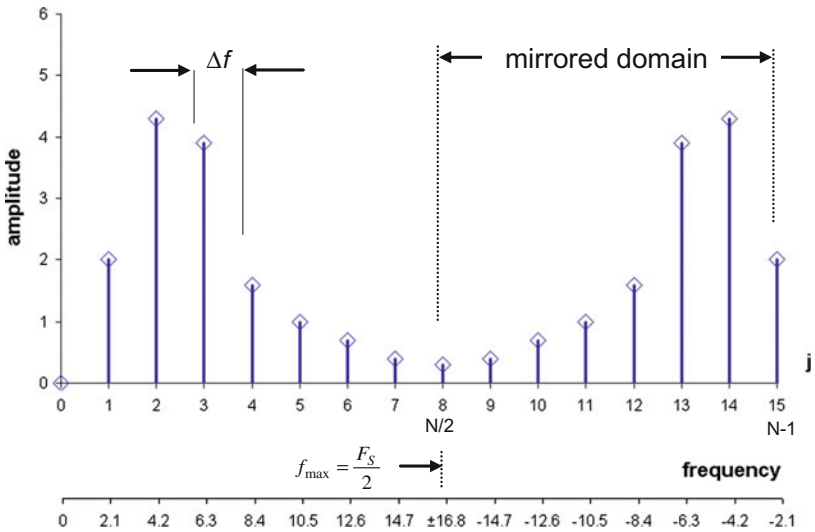


Fig. 3 Amplitude spectrum of a sampled signal

$$f_{Ny} = \frac{N}{2}\Delta f \approx \frac{1}{2\tau} = \frac{F_S}{2}, \tag{15}$$

called the Nyquist frequency. An analysis regarding the influence of the nature of N on the spectral lines distribution and the maximum achievable frequency $f_{\max}$ is worth of attention. It shows that the estimated number of samples $N/2$ which should be involved in the spectral analysis is a row approximation. A precise evaluation of $f_{\max}$ is made in Table 1, constructed using the property stated in Eq. (14).

If N is an even number, the equivalent frequencies are on the lines

$$N - j \equiv N - j + N = -j \tag{16}$$

whereas, if N is an even number, the equivalent frequencies are on the lines

$$\frac{N-1}{2} + j \equiv \frac{N-1}{2} + j - N = j - \frac{N-1}{2} \tag{17}$$

The switch from positive to negative frequencies takes place for the first frequency whose value would be equal to or larger than the Nyquist frequency. The index of these frequencies and the frequencies itself are marked in dark background in Table 1. Note that the index j for which this happens, if even number of samples are chosen, is actually $(N/2) + 1$. For an odd number of samples, the index for which $f_{\max}$ is achieved is $(N + 1)/2$.

A more powerful tool for frequency evaluation is the power spectrum (PS), defined as

$$P[f_j] = \frac{1}{N}\left|\sum_{k=0}^{N-1} x[k]e^{-i2\pi f_j t}\right|^2 = \frac{1}{N}\left|X[f_j]\right|^2 \tag{18}$$

Table 1 Spectral line distribution and the corresponding frequencies

Index number	N is an even number		N is an odd number	
	Index value j	Frequency f_j	Index value j	Frequency f_j
1	0	DC	0	DC
2	1	Δf	1	Δf
3	2	$2\Delta f$	2	$2\Delta f$
…	…	…	…	…
$j + 1$	$\frac{N}{2}$	$\pm\frac{N}{2}\Delta f = f_{\max}$	$\frac{N-1}{2}$	$\frac{N-1}{2}\Delta f = f_{\max}$
…	…	…	$\frac{N-1}{2}+1$	$-\frac{N-1}{2}\Delta f$
…	…	…	…	…
$N-1$	$N-2$	$-2\Delta f$	$N-2$	$-2\Delta f$
N	$N-1$	$-\Delta f$	$N-1$	$-\Delta f$

which is typically used to examine the various frequency components of a signal. The representation of all frequency-amplitude pairs for all spectral lines characterize the signal in the frequency domain and is the so-called periodogram. Note that, in case of PS the spectral lines are identically distributed as in the case of DFT. Further we will use the PS.

The main disadvantage of these frequency estimators is the weak frequency resolution, especially in the case of short-time signals, as in damage detection occurs. This will be demonstrated by applying the PS to a sequence representing a sinusoidal signal with frequency $f = 4$ Hz and amplitude $A = 1$. Two cases are exemplified. In the first case, the sequence has a length $T_{S1} = 1$ s, therefore containing an integer number of periods, as shown in Fig. 4a. The PS of this signal, plotted with a red line in Fig. 5, precisely indicate the frequency of 4 Hz at the 5-th spectral line. The amplitude is also correct indicated, the value 0.5 being expected since the energy is distributed into two spectral lines. Note that just one of them is visible in Fig. 5, because the spectrum is a single-sided representation.

In the second case, the signal sequence has the length $T_{S2} = 1.1$ s. As shown in Fig. 4b, it does not contain an integer number of periods. Its PS is plotted with a blue line in Fig. 5. Even if the spectral lines are denser, the frequency is not accurately evaluated, while we found a peak at 3.6036 Hz instead of 4 Hz. For this frequency, the amplitude 0.2748 is obtained in stead of 0.5 as expected. Thus, neither the frequency nor the amplitude is correct indicated.

This happens, for the second example, because the energy is dispersed into "phantom" frequencies while no spectral line coincides with the real frequency. The phenomenon, known as spectral leakage, causes also an apparent amplitude decrease because the energy is divided between more spectral lines [12]. The estimated frequency can differ from the real frequency value with at most a half of the frequency resolution, which is $\varepsilon_{max} = \Delta f/2$. This error cannot be predicted, because it depends on the signals period $T = 1/f$, which is obviously not known before the measurement takes place. Several techniques to improve the frequency readability are known. We present in the next sections three simple methods; they will be tested for a sinusoidal signal against an original method proposed by the authors.

4 Simple Methods to Improve the Frequency Readability

The simplest attempt to improve the frequency readability consists in increasing the observation time T_S. The increased signal length has as consequence the decrease of the width between tow consecutive spectral lines, i.e. a finer frequency resolution Δf. Figure 6 presents the PS for the sinusoid described in the previous section, having the frequency $f = 4$ Hz and the length $T_{S1} = 1.1$ s. The signal was generated by $N = 110$ samples, with a sampling frequency rate $F_S = 100$ Hz. The PS indicates a false frequency, $f_1 = 3.6036$ Hz, the resulted error being $\varepsilon_1 = 0.3964$ Hz.

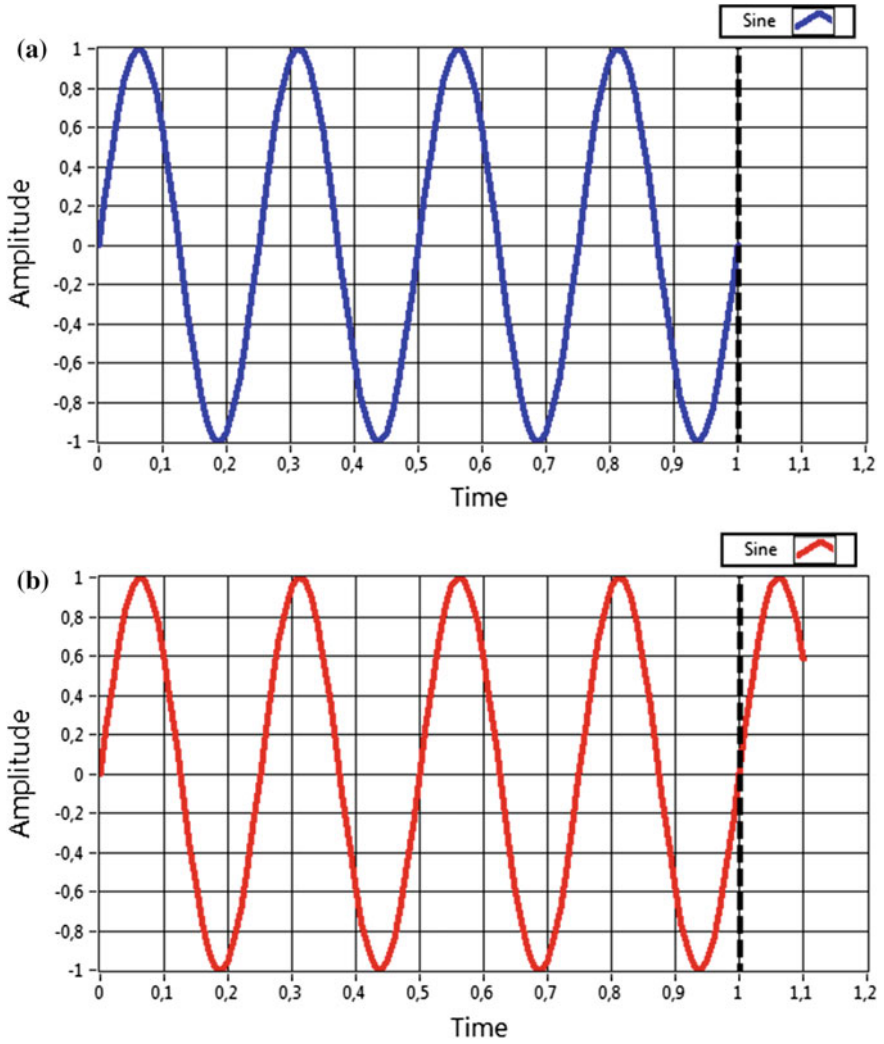


Fig. 4 Sinusoid with integer number of periods (**a**) and non-integer number of periods (**b**)

For the signals with increased observation time, $T_{S2} = 2.2$ s and $T_{S3} = 3.1$ s respectively, the frequency resolution Δf decrease two respective three times, in accordance to Eq. (11). The signals, presented in Fig. 7a are generated at $N_2 = 221$ respectively $N_3 = 311$ samples. One can identify in Fig. 6 the frequencies $f_2 = 4.0724$ Hz respectively $f_3 = 3.8585$ Hz; both of them are more precisely as that found for T_{S1}. The errors are thus reduced to $\varepsilon_2 = 0.0724$ Hz respectively $\varepsilon_3 = 0.1415$ Hz. This experiment shows that a finer frequency resolution leads to lower maximum possible errors, but it does not guarantee an improvement in the frequency estimation.

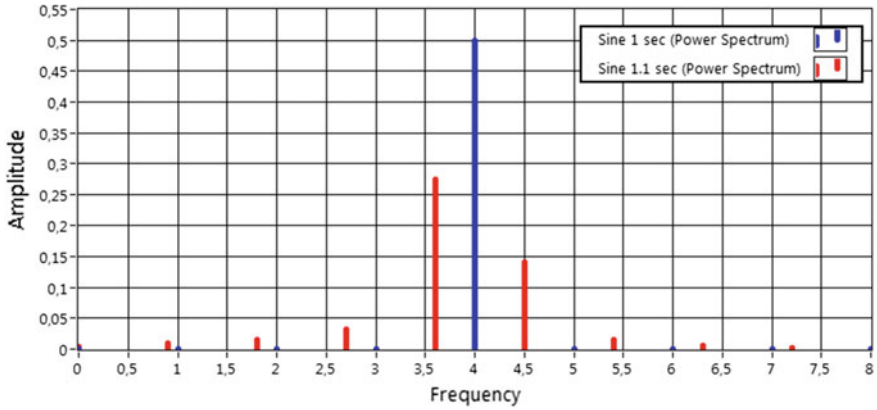


Fig. 5 Power spectrum of the two signals

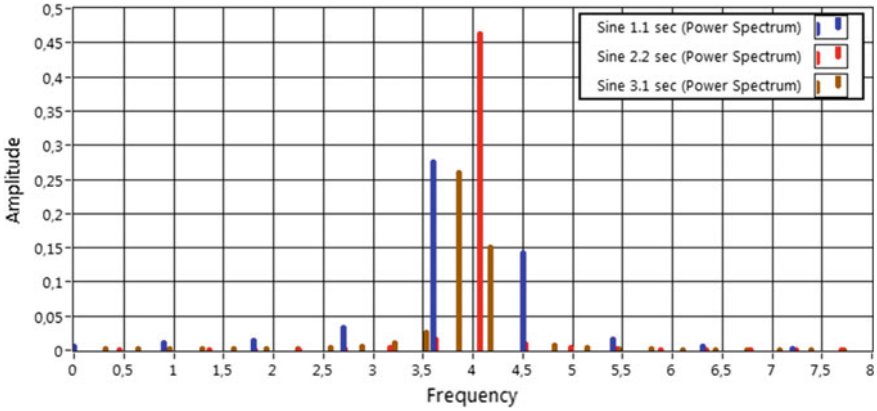


Fig. 6 Reduced distortion effect

If increasing the observation time cannot be done, for instance in the cases of signals acquired from free damped vibrations, the signal length can be artificially extended by adding a number of samples for which the amplitude is null [13, 14]. The procedure, illustrated in Fig. 7b, is known as "zero-padding". The exemplified signal has the length $T_{S3-ZP} = 3.1$ s, and it takes the same frequency resolution $\Delta f_3 = 0.322851$ Hz as the sinusoid with the similar time length (see Fig. 8). The disadvantage of this procedure is the dramatic decreasing of the amplitudes, which can make it not operative. Table 2 presents the settings of the three simulated signals and the PS output in terms of frequencies and amplitudes.

Signal windowing is considered an alternative method which improves the frequency readability. When performing Fourier or spectral analysis on finite-length data that contain non-integer number of cycles, the application of a window

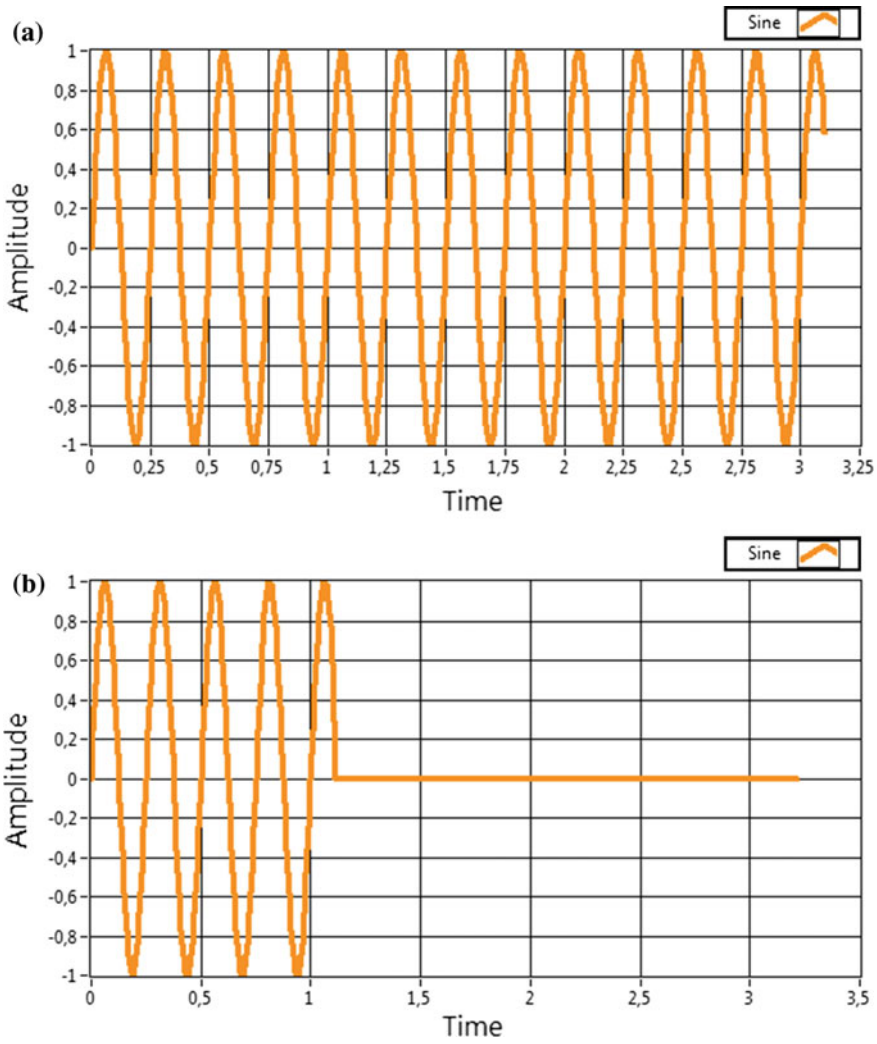


Fig. 7 The signal with extent length due to observation time increasing (**a**) and zero-padding (**b**)

smoothes the signal ends, making the amplitude to vary gradually towards zero at the ends [15, 16]. This minimizes the discontinuities of the finite-time signal edges, reducing spectral leakage.

Firstly, the effect of windowing is analyzed for a sine with the frequency $f = 4$ Hz and the amplitude $A = 1$. The signal is generated with $N = 321$ samples by a sampling frequency $F_S = 100$ Hz, resulting in a signal time length $T_S = 3.2$ s. The PS for the original signal is plotted in Fig. 9. In addition, three other spectra are plotted for the signal for which three types of windows were applied: Hamming, Blackman-Harris and Flat top. One can observe in Fig. 9 that windowing does not

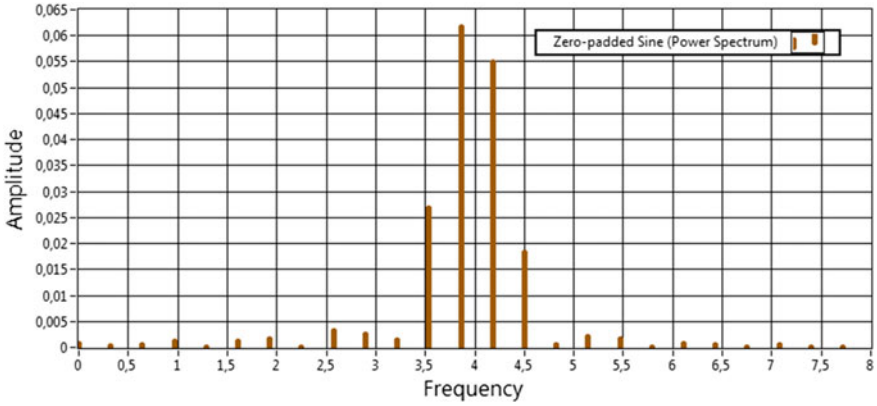


Fig. 8 Increasing the frequency resolution through zero-padding

Table 2 Settings for the frequency evaluation and the achieved results

Signal name	Color	Signal settings				PS results		
		T_S [s]	N	F_S	Δf [Hz]	f [Hz]	A	ε [Hz]
Sine 1	Blue	1.1	111	100	0.909091	3.6036	0.2748	0.3964
Sine 2	Red	2.2	221	100	0.454545	4.0724	0.4674	−0.0724
Sine 3	Brown	3.1	311	100	0.322851	3.8585	0.26	0.1415
Zero-padded	Brown	1.1 + 2	101 + 200	100	0.322851	3.8585	0.0625	0.1415

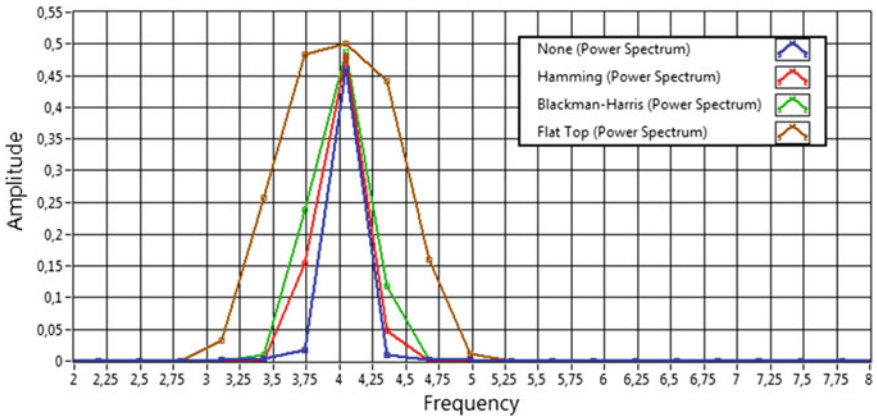


Fig. 9 PS of a sine in the absence and the presence of windowing

improve the frequency readability since the spectral lines maintain their location (Δf is unchanged). Moreover, for a single-tone signal, the amplitudes indicated at the spectral lines next to that corresponding to the pick amplitude have increased values.

A second test regarding the signal windowing is performed on a two-tone signal, composed by two sinusoids with the frequencies $f_1 = 4$ Hz respectively $f_2 = 7$ Hz, and amplitudes $A_1 = 1$ respectively $A_2 = 0.0025$. The signals were generated using $N = 325$ samples, with a sampling frequency $F_S = 100$ Hz, resulting in a signal time length $T_S \approx 3.2$ s. The PS in linear representation, presented in Fig. 10a, is not able to indicate both frequencies. In contrary, the representation of the results in dB plotted in Fig. 10b highlights the existence of a second frequency at 7 Hz, even if this frequency component has for very small amplitude. However, the frequency values cannot be accurately extracted.

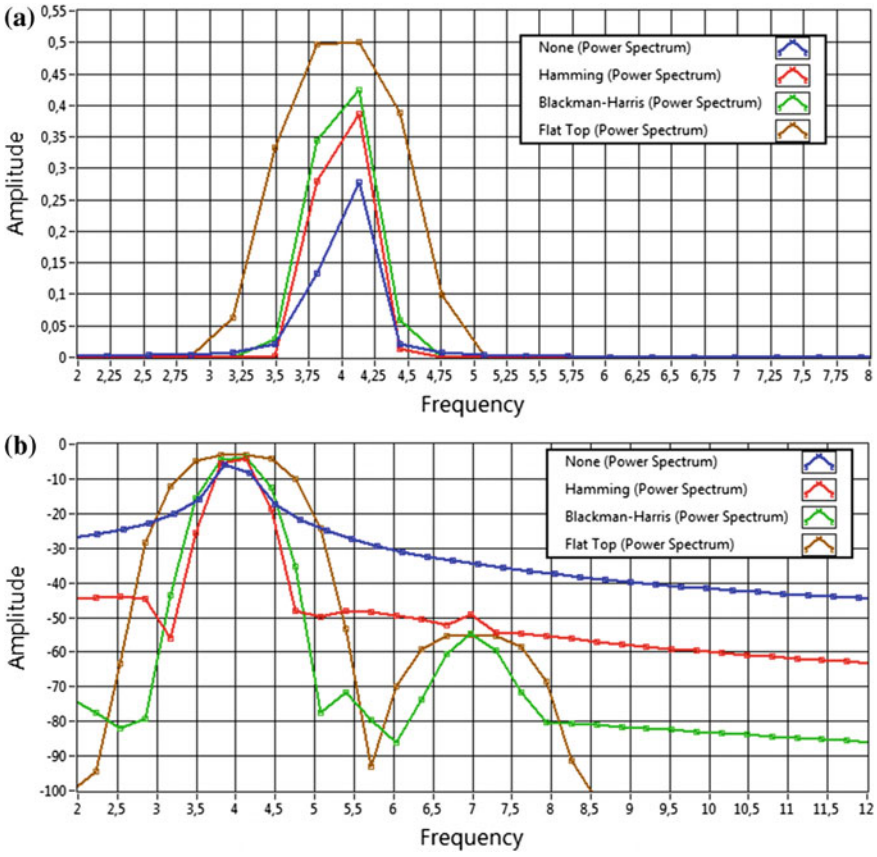


Fig. 10 Representation of the PS for a two-tome signal in the absence and the presence of windowing: **a** linear, and **b** dB representation

As a conclusion, the previously presented methods can be used to reduce the possible error range by decreasing the distance between the spectral lines (extension of the observation time or zero-padding), or to indicate the existence of closely located frequencies (windowing). But, for short-time signals commonly used in damage detection, as free damped vibrations are, the achieved precision is insufficient, so that other methods have to be investigated.

A method for which an inter-line frequency is found relies on plotting a curve best fitting to three consecutive points obtained in the PS: B(f_{j-1}, A_{j-1}), C(f_j, A_j) and D(f_{j+1}, A_{j+1}). Among them, C is always a local maximum in the PS, as shown in Fig. 11. Thus, the evaluated frequency is not directly linked with the spectral line position in the periodogram. The frequency f_{corr} considered as correct is correlated with the amplitude $A_{\max}$ found as the curve's pick value. The fractional correction term is $\delta = f_{\text{corr}} - f_j$.

There are several functions used to evaluate the peak location, and therefore the corrected frequency in the periodogram. In [17–19] parabola approximation for peak determination is used, so the correction term is given by

$$\delta = \frac{A_{j+1} - A_{j-1}}{4A_j - 2A_{j-1} - 2A_{j+1}} \tag{19}$$

For exemplification of the method, a sine signal with a generated frequency $f = 4$ Hz was analysis. As shown, the frequency resolution $\Delta f = 1/T_S$ dependents on the signal time length T_S. The aim of this experiment is to highlight the influence of the observation time T_S upon the method's precision. Therefore, signals with time lengths between $T_{\min} = 0.65$ s to $T_{\max} = 1.35$ s have been created. Table 3 presents 12 simulation cases, reflecting the phenomenon for a signal length of 1 cycle, around the central observation time of $T_S = 1$ s. Obviously, different frequency resolutions are achieved for the different time lengths, and the frequency-amplitude

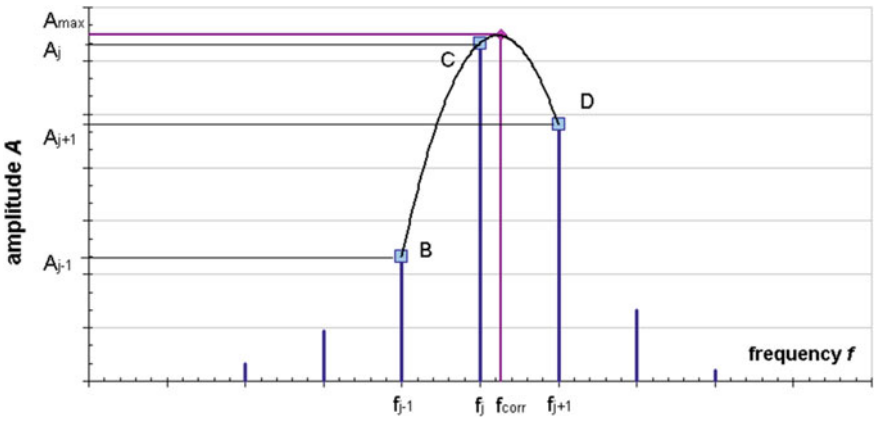


Fig. 11 The interpolation of three points from PS

pairs result in consequence. The standard frequency evaluation is preformed by involving the PS for each observation time. Three of the amplitudes are indicated for all simulation cases; always the pick value is taken in central position. For this value the frequency estimated by the standard method is also provided. Figure 12 shows the pick amplitude for all analyzed cases.

The fractional correction term δ is derived using Eq. (19) from the sequence of amplitudes specific for each time length. This parameter is used to correct the read frequency. Figure 13 presents a comparison between the frequencies attained by standard evaluation and those derived with the correction term. It results that there are time lengths for which the correction improves the results, but the frequencies are still not exact estimated.

Two specific domains, located around the two types of characteristic points, are observable in Fig. 13. The first characteristic points are defined by the time lengths for which the signal achieves integer number of cycles (e.g. T_S = 1 s). Thus, the first domain is symmetrically located around these points, for instance the T_S in the range 0.95–1.05 s. Here, the frequencies are similarly estimated, irrespective to the evaluation method. The closer the T_S to the domain center, the higher the precision in evaluating frequencies is, because the position of the spectral line becomes closer to the real frequency. If the T_S match an integer number of cycles, e.g. 1 s, the frequency is perfectly estimated.

The second type of characteristic points are centered between the first one, so that the time lengths past with one half the integer number of cycles. For the signal with frequency f = 4 Hz, on of these point is T_S = 1.125 s. Around these points a second domain exists, for instance the T_S in the range 1.05–1.2 s. Having a look onto Fig. 13 one can observe that, for these domains, the interpolation method improves the frequency readability, but even so, the errors are more important as that achieved in the first domain.

Table 3 Frequency achieved by standard evaluation and after correction with the term δ

T_S [s]	Δf [Hz]	f_{read} [Hz]	A_{j-1} [–]	A_j [–]	A_{j+1} [–]	δ [Hz]	f_{corr} [–]	error [%]
1.12	0.892857	3.57143	0.03298	0.24431	0.16687	0.23184	3.80327	−4.92
1.1	0.909091	3.63636	0.03136	0.30907	0.11468	0.08825	3.72461	−6.88
1.08	0.925926	3.70370	0.02457	0.36470	0.07369	0.03892	3.74262	−6.43
1.06	0.943396	3.77358	0.01555	0.41093	0.04182	0.01718	3.79077	−5.23
1.04	0.961538	3.84615	0.00744	0.44977	0.01844	0.00629	3.85245	−3.69
1.02	0.980392	3.92157	0.00203	0.48110	0.00439	0.00124	3.92281	−1.93
1	1	4	0	0.5	0	0	4	0
0.98	1.020408	4.08163	0.00291	0.49826	0.00329	−0.00019	4.08183	2.05
0.96	1.041667	4.16667	0.01473	0.46947	0.01011	−0.00253	4.16414	4.10
0.94	1.06383	4.25532	0.04106	0.41407	0.01601	−0.01624	4.23908	5.98
0.92	1.086957	4.34783	0.08572	0.34072	0.01851	−0.05822	4.28960	7.24
0.9	1.111111	4.44444	0.14709	0.26275	0.01783	−0.17925	4.26520	6.63

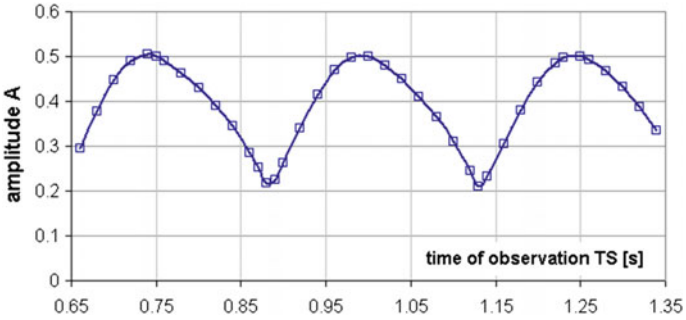


Fig. 12 PS pick values A_j versus observation time T_S

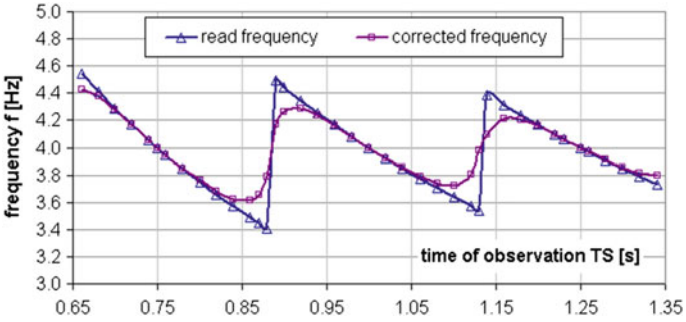


Fig. 13 Estimated frequency values f_{read} and f_{corr} versus observation time T_S

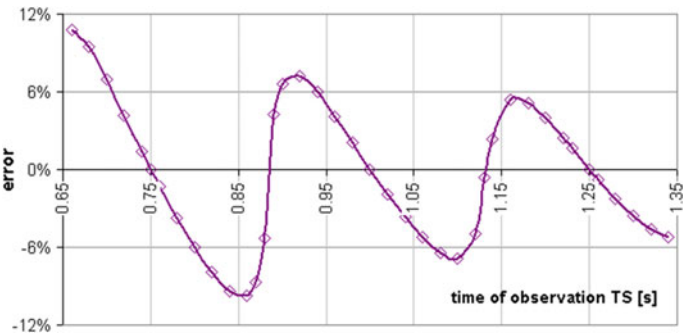


Fig. 14 Errors attained for the estimated frequencies f_{read} and f_{corr}

The errors resulted from frequency evaluation, if the correction term is applied, are plotted in Fig. 14. One can observe that the error varies cyclically, depending on the central point C position in the PS. It should also be noted that the maximum expected error decrease with increasing of the observation time.

Nor of the methods above presented satisfy the requests of precision in frequency evaluation for damage detection processes [20, 21]. Early, effective damage detection imposes accurate frequency estimation and, in addition, the possibility to observe minor changes in the modal parameters if physical or geometrical parameters of a structure are lightly affected [22]. This implies the use of simple but performant signal processing algorithms, which are capable to provide power spectra with very dense spectral lines, even for short-time signals. The next section describes an original signal post-processing algorithm who permits a dramatically improvement of the frequency readability.

5 Description and Implementation of the Iterative Algorithm

The algorithm is step-described, in order to clearly get the right idea about how it works [23]. In the main, it acts as an iterative loop, decreasing a certain number of samples from the original acquired signal, converting each achieved signal in a frequency spectrum and overlapping the spectra in order to realize a higher resolution spectrum. In the following, this spectrum will be assumed as overlapped-spectrum (OS). The overlapped-spectrum instead of one certain spectral line (SL) of significant amplitude shows a lobe shape formed by a number of bins equal to the number of iteration.

Figure 15 aims only to illustrate the SLs distribution in the case of OS. Therefore, all the amplitudes of SLs are equal to the unit. Three superposed spectra (S_1, S_2, S_3) achieved after applying the algorithmic post-processing procedure are supposed. The frequency resolutions are in the following relation $\Delta f_1 < \Delta f_2 < \Delta f_3$, corresponding to the signal lengths $T_1 > T_2 > T_3$.

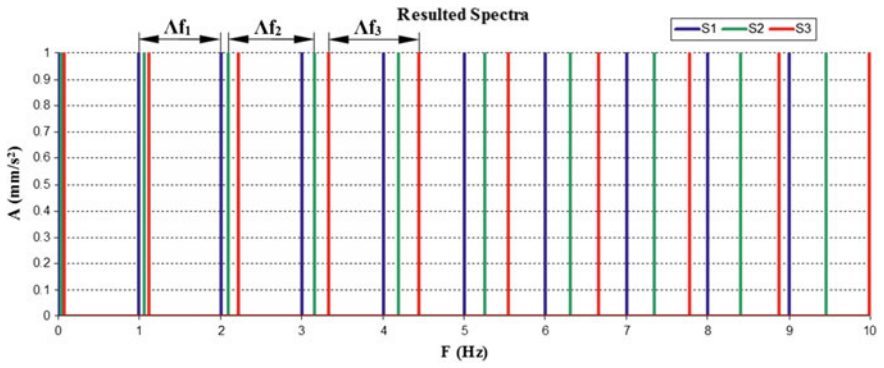


Fig. 15 SLs distribution for the OS resulted by superposing three spectra with different frequency resolutions: Δf_1 = 1 Hz, Δf_2 = 1.05 Hz and Δf_3 = 1.11 Hz

In Fig. 15 can be observed that even if frequency resolution decreases, by superposing those spectra the distance between SLs has been decreased. This means that the frequency resolution of the resulted OS significantly increases in comparison to one of the individual spectra. Also, even if the bins in the overlapped spectrum are not equidistant, the higher density greatly increases and thus the probability as one of the SLs to fit more accurately the desired frequency.

For a faster understanding, the algorithm working principle can be resumed to the following steps:

- it takeover the acquired data, corresponding to the vibration signal in the time-domain;
- a certain number of samples are iteratively decreased, this could be assumed as a rectangular windowing applied to the signal;
- a Power Spectrum analysis is performed for all truncated signals, in this way individual spectra are achieved;
- in the end, all the achieved spectra are overlapped to obtain an OS of much higher resolution.

It has to be taken into account the algorithm steps must be repeated for each desired frequency f. The desired frequency value is in relation with the decreased sample's number N_S via its corresponding period T, acquisition period T_S and sample's number of the original signal N:

$$N_S = \left(1 - \frac{1.3}{n}\right) \cdot N, \quad n = \frac{T_S}{T} \tag{20}$$

where $n > 1.3$.

Also, after algorithm decreased a number of samples N_S equivalent to the time length T_D, the variation of frequency resolution can be appreciated:

$$\Delta f^* = \Delta f_D - \Delta f \tag{21}$$

where Δf is the initial resolution, corresponding to the T_S, and the resolution after the signal portion is cropped out:

$$\Delta f_D = \frac{1}{T_S - T_D} \tag{22}$$

As pre-conclusion, it should be said the iterative algorithm offers the possibility to more accurately evaluate the frequency insofar a denser spectrum is achieved.

As logic structure, the iterative algorithm has been implemented in a numerical form by the help of LabView software. In the main, it consists of a conditional loop, having one input and one output (Single Input—Single Output SISO). Data file corresponding to the acquired vibration signal is applied to the input, at the output accomplishing the OS in a graphical view.

The "Block Diagram" (BD) window of LabView ensures the placement and interconnectivity of functional blocks. As example, instead of real vibration signal, the input data can be provided from an auto-generated signal by using the functional block "Simulate Signal" (SS), where the waveform, amplitude, frequency, length and number of samples were previously set. The more important part of algorithm block diagram (ABD) is the logic structure placed in the conditional loop, where "Extract Portion of Signal" (EPS) and "Spectral Measurements" (SM) blocks perform signal cropping and spectral analysis. At the end of iteration, all accomplished data are sending to the "Graph" (G) block named Display Resulted Spectra (D.R.S.). Data are presented as graphical OS in a different window "Front Panel" (FP). The parameters chosen to be set or visualize are shown in FP as well.

Figure 16 presents the BD image that is meant to comprehensively clear up all the aspects associated with the algorithm implementation. In the image can be also seen two bordered areas, which specify the alternatively way to input the signal

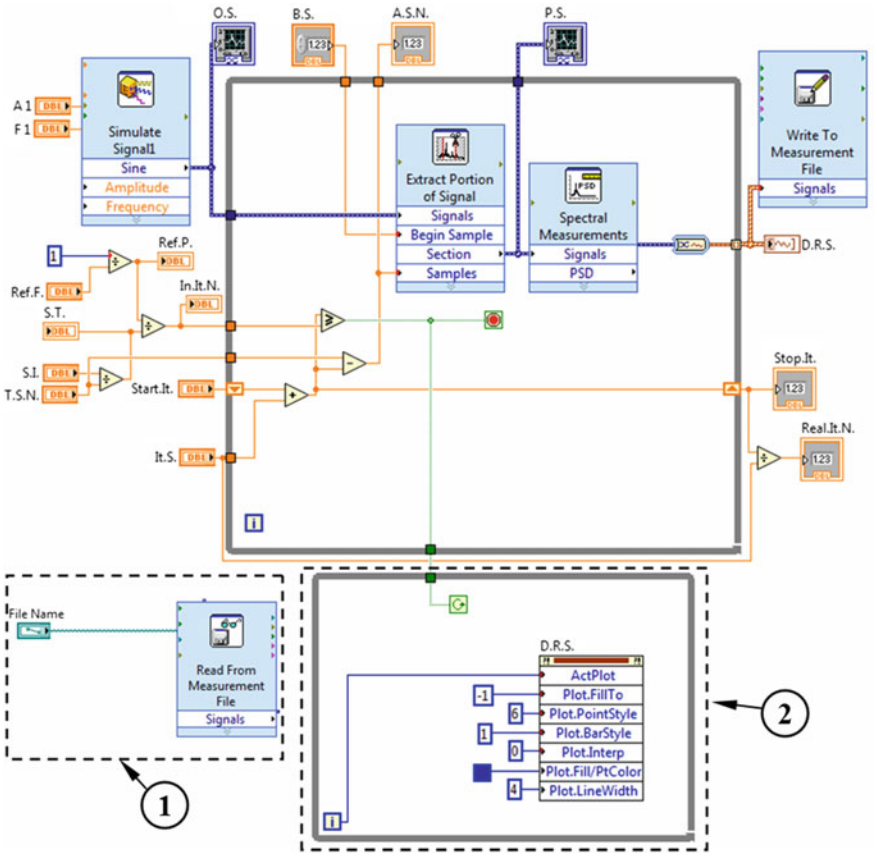


Fig. 16 Software implementation of iterative algorithm

(1) of analysis and graphically display the results (2). The algorithm input can be provided from one or more SSs, or from a data file ("File Name") via the "Read From Measurements" (RFM) block. Data file must contain an Analog to Digital Conversion (ADC) data stream of a real vibration signal. The (2) block performs modifications on graphical view of the results, in order to have more options for displaying the frequency bins. Same color for all frequency bins not allows the individual spectra to be distinguished; it only shows the resolution increasing. Different color for each spectrum highlights the participation of any spectra to the resolution of OS.

The abbreviated forms of parameters included in the Fig. 16 are explained Table 4. Also, for each of them is presented the sense of consideration (as input or output).

Above parameters should not be considered only for simple signal cropping and iteration, but also to create many possibilities of data and results management from whatever section of the algorithmic structure. Therefore, considering important the way in which the iterative algorithm performs in applications, to get suitable information and more details some examples are presented in the next chapter.

Table 4 Parameter's meaning

Parameter	Sense	Detail
A1	Input	Signal amplitude
F1	Input	Signal frequency
Ref.F.	Input	Reference frequency—expected to reach by analysis
Ref.P.	Output	Reference period—corresponds to Ref.F.
T.S.N.	Input	Total samples number—all signal samples considered in test
S.I.	Input	Sampling interval—whole signal length in time
S.T.	Output	Sampling time
B.S.	Input	Begin sample—first sample from the signal left-side
A.S.N.	Output	All samples number—remaining samples after last iteration is done
Start.It.	Input	Start iteration—sample whence algorithm begins the iterative cropping
It.S.	Input	Iteration step—number of samples cropped at iteration
Stop.It.	Output	Stop iteration—number of samples cropped at the iteration end
In.It.N.	Output	Initial iteration number—resulted by computing
Real.It.N.	Output	Real iteration number—resulted by applying It.S. to In.It.N. and rounding
O.S.	Output	Original signal—graphically displayed
P.S.	Output	Portioned signal—graphically displayed when the iteration ends
D.R.S.	Output	Display resulted data—data displayed as overlapped-spectrum

6 Testing the Algorithm Efficiency

These examples were considered to check the algorithm efficiency. In the first example the algorithm performs the analysis of one auto-generated signal at a precise value of 9.753 Hz. Set values for the involved parameters are given into the Table 5. The dark background marks the parameter auto-computed values after the algorithm starts.

In order to have a clear view on the cropped portion of the original signal, Fig. 17a shows entire signal and Fig. 17b the remained portion, after algorithm decreased a number of samples equivalent to the period T_D that corresponds to 9.5 Hz.

Figure 18 presents the main lobe formed around to 9.753 Hz and a zoomed area of lobe's peak. All the spectra of OS are displayed in blue. In the zoomed area it is can be observed that a certain frequency bin precisely fits the 9.753 Hz value.

The second example aims the analysis of a real vibration signal. The input signal was previously acquired using the test stand depicted in the Fig. 19. In experiment a cantilever beam fixed to the left-side has been involved.

The cantilever beam has the geometric and mechanical characteristics given in Table 6.

The meaning of the parameters involved in Table 6 is: l—length of a beam, w—width of the beam, t—thickness of the beam, I—moment of inertia, ρ—mass density, E—Young modulus and v—Poisson's ratio. Left-end of the beam was fixed in a milling machine vise.

The acquisition hardware equipment consists in: computer, compact chassis NI cDAQ-9172, signals acquisition module NI 9234 and accelerometer Kistler 8772. The accelerometer is mounted near by the free end of the beam.

Table 5 Main parameters involved in the algorithm

S.I. [s]	T.S.N.	S.T. [s]	Ref.F. [Hz]	Start It.	It.S.	Real It.N.
1	10,000	10^{-4}	9.5	0	4	250

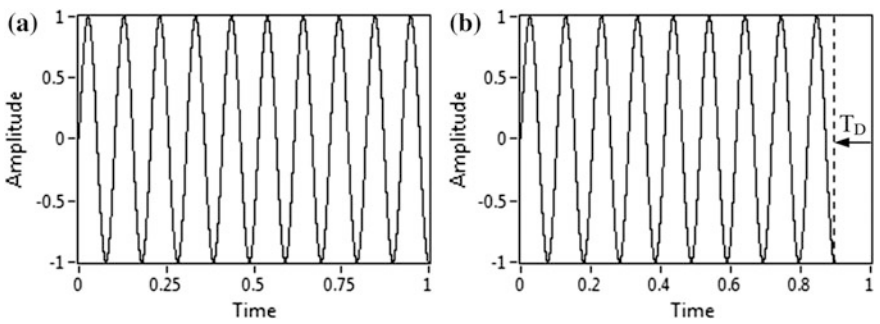


Fig. 17 Entire signal (**a**) and remained portion after T_D segment was cropped (**b**)

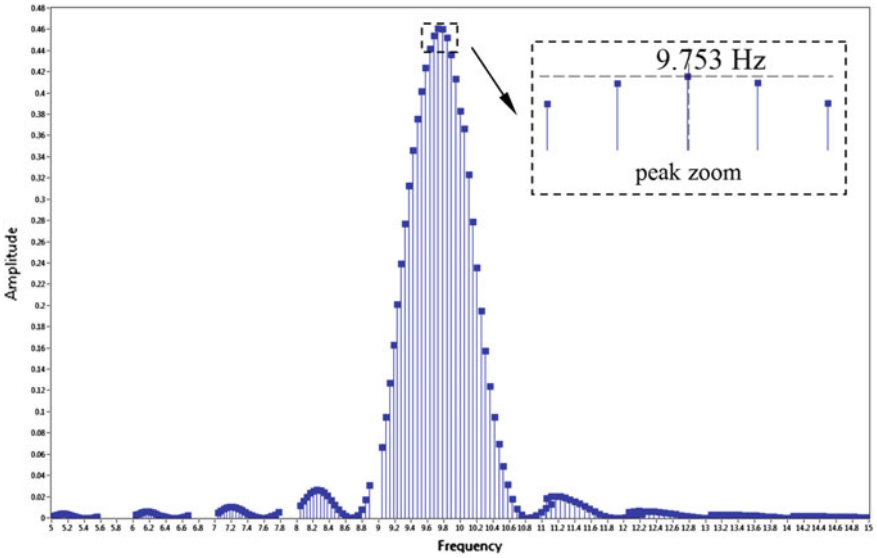


Fig. 18 Main lobe formed at 9.753 Hz and the peak zoomed area

Fig. 19 View of the experimental stand

Table 6 Geometric and mechanical characteristics of the cantilever beam

l [mm]	w [mm]	p [mm]	t [mm]	I [m^4]	ρ [kg/m^3]	E [N/m^2]	v [–]
1000	50	410	5	520.833×10^{-12}	7850	2.0×10^{11}	0.3

The algorithm has to be tuned up for each desired frequency, as explained above, and so that two examples of setup, for the frequencies of first two vibration modes, are given in Table 7.

Figure 20 presents the spectrum adjusted area in order to display as clearly as possible the above mentioned two frequencies, with a zoom on the peak of the first frequency. Here, the different color for each spectrum of OS can be remarked.

By the algorithm setup the achievement of the first five weak-axis bending vibration modes has been performed, each mode has been analyzed for five times and the average value was kept as final. The values are given in the Table 8.

Table 7 Parameter values involved in the algorithm setup

S.I. [s]	T.S.N.	S.T. [s]	Ref.F. [Hz]	Start It.	It.S.	Real It.N.
10	50,000	2×10^{-4}	9.5	100	10	57
10	50,000	2×10^{-4}	65	0	1	77

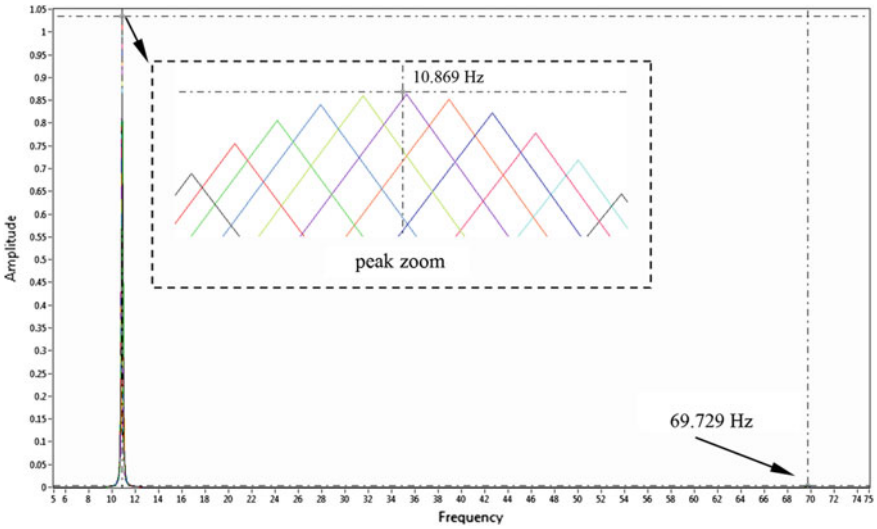


Fig. 20 Frequencies corresponding to first two vibration modes

Table 8 First five natural frequencies achieved for the undamaged cantilever beam

Mode number	Test 1 [Hz]	Test 2 [Hz]	Test 3 [Hz]	Test 4 [Hz]	Test 5 [Hz]	*A* [Hz]
1	10.86955	10.86896	10.8686	10.87029	10.86977	10.86943
2	69.74296	69.73753	69.73018	69.74354	69.742	69.73924
3	195.2166	195.1993	195.1931	195.21	195.2072	195.2052
4	383.1953	383.0578	383.1829	383.2283	383.0092	383.1347
5	633.5744	634.2199	633.7255	633.3499	632.7726	633.5285

Analyzing these values little differences between the values of frequencies for each mode are observed, they being less than 0.97 Hz in absolute and 0.6% in percentage. The differences can be explained as occurring due to the initial phase of the different modes and the relative values of the acceleration amplitudes.

As conclusion of this chapter, performing many analyses on cantilever beams with different dimensions, having a significant variation of damages, as location and severity, method confirms its effectiveness in improving the readability of frequencies, especially in the case of short time signals [24].

7 Conclusions

In this paper, a signal post-processing method for accurate identification of the natural frequencies was comprehensively presented. The method is based on an iterative algorithm, which performs the step-cropping of the acquired vibration signal, analyzes the truncated signal at each step and realizes a much denser overlapped-spectrum. This overlapped-spectrum offers the possibility to identify natural frequencies with higher precision.

The algorithm working way has been presented and exemplified for a generated signal and same real vibration signals. In all cases, more accuracy in frequency identification was accomplished. Therefore, the method is strongly recommended in case of rapid vibration damping and small cracks. The method work-limits depend only on the precision of signal acquisition and sampling frequency value, mainly for short time signals.

Acknowledgements The work has been funded by the Sectoral Operational Programme Human Resources Development 2007–2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/132395.

References

1. Doebling S.W., Farrar C.R., Prime M.B., “A summary review of vibration based damage identification methods,” Shock Vibration Digest, 1998, 30(2): 91–105.
2. Salawu O.S., “Detection of structural damage through changes in frequency: a review,” Engineering Structures, 1997, 19(9): 18–723.
3. Morassi A., Vestroni F., “Dynamic methods for damage detection in structures,” CISM Courses and Lectures, Vol. 499, Springer Wien New York, 2008.
4. Gillich G.R., Praisach Z.I., “Modal identification and damage detection in beam-like structures using the power spectrum and time-frequency analysis,” Signal Processing 2014, 96 (PART A): 29–44.
5. Friswell M.I., “Damage identification using inverse methods,” Phil. Trans. R. Soc. 2007, A 365: 393–410.

6. Gillich G.R., Praisach Z.I., Wahab M.A., Vasile O., "Localization of transversal cracks in sandwich beams and evaluation of their severity," Shock and Vibration 2014, Article Number: 607125.
7. Rytter A., Vibration Based Inspection of Civil Engineering Structures, Ph.D. Thesis, Aalborg University, Denmark, 1993.
8. Hutin C., "Modal analysis using appropriated excitation techniques," Sound and Vibration 2000, 34(10): 18–25.
9. Gillich G.R., Praisach Z.I., Iavornic C.M., "Reliable method to detect and assess damages in beams based on frequency changes," Proceedings of the ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference 2012, Vol. 1: 129–137.
10. Maia N.M.M, Silva J.M.M., Theoretical and Experimental Modal Analysis, Research Studies Press, Taunton, Somerset, UK, 1997.
11. National Instruments, LabVIEW Analysis Concepts, March 2004 Edition, Part Number 370192C-01.
12. Minda A.A., Gillich N., Mituletu I.C., Ntakpe J.L., Manescu T., Negru I., "Accurate frequency evaluation of vibration signals by multi-windowing analysis," Applied Mechanics and Materials 2015, vol. 801: 328–332.
13. Chioncel C.P, Gillich N., Tirian G.O., Ntakpe J.L., "Limits of the discrete Fourier transform in exact identifying of the vibrations frequency," Romanian Journal of Acoustics & Vibration 2015, 12(1): 16–19.
14. Donciu C., Temneanu M., "An alternative method to zero-padded DFT," Measurement 2015, 70: 14–20.
15. Andria G.; Savino M., Trotta A., "Windows and interpolation algorithms to improve electrical measurement accuracy," IEEE Transactions on Instrumentation and Measurement, 1989, 38 (4): 856–863.
16. Abed S.T., Dallalbashi Z.E., Taha F.A., "Studying the effect of window type on power spectrum based on MatLab," Tikrit J. Eng. Sciences 2012, 19(2): 63–70.
17. Jacobsen E., Kootsookos P., "Fast, accurate frequency estimators," IEEE Signal Processing Magazine 2007, 123: 123–125.
18. Candan Ç., "A method for fine resolution frequency estimation from three DFT samples," IEEE Signal Processing Letters, 2011, 18(6): 351–354.
19. Voglewede P., "Parabola approximation for peak determination," Global DSP Mag. 2004, 3 (5): 13–17.
20. Gillich G.R., Praisach Z.I., "Detection and quantitative assessment of damages in beam structures using frequency and stiffness changes," Key Eng. Mat. 2013, 569: 1013–1020.
21. Gillich G.R., Birdeanu E.D., Gillich N., Amariei D., Iancu V., Jurcau C.S., "Detection of damages in simple elements," Proceedings of the International DAAAM Symposium 2009, pp. 623–624.
22. Gillich G.R., Maia N.M.M., Mituletu I.C., Praisach Z.I., Tufoi M., Negru I., "Early structural damage assessment by using an improved frequency evaluation algorithm," Latin American Journal of Solids and Structures, 2015, 12(12): 2311–2329.
23. Mituletu I.C., Gillich N., Nitescu C.N., Chioncel C.P., "A multi-resolution based method to precise identify the natural frequencies of beams with application in damage detection," Journal of Physics: Conference Series 2015, 628(1), 012020.
24. Gillich G.R., Mituletu I.C., Negru I., Tufoi M., Iancu V., Muntean F., "A method to enhance frequency readability for early damage detection," Journal of Vibrational Engineering and Technologies, 2015, 3(5), 1: 637–652.

Holobalancing Method and Its Improvement by Reselection of Balancing Object

Yuhe Liao and Liangsheng Qu

Abstract Based on the idea of multi-sensor information fusion, one of the main problem—insufficient utilization of rotor vibration information—existing in the traditional rotor balancing methods is solved. By integration of all the amplitude, frequency and phase information, the Holobalancing method can help to correct the rotor balancing state more accurately and efficiently than other traditional methods. The field balancing capability has been greatly improved therefore. Since the Holobalancing method truly considers the characteristics of system support stiffness anisotropy, the unreasonable isotropic assumption adopted in traditional balancing methods is no longer required therefore. The balancing result of the Holobalancing method is more reliable and fewer number of trial runs is needed. Recently, the Holobalancing method is further improved by reselection of balancing object. With the Initial Phase Vector (IPV) being replaced by its forward precession component (IPV+), the impact of probe orientation on the balancing analysis and calculation is completely eliminated and the computational procedure is greatly simplified without sacrificing the balancing accuracy. The experiments and field application cases verify the effectiveness and reliability of this method.

1 Introduction

The Holobalancing method is developed on the basis of the Holospectrum technique [1, 2]. It is the application of multi-sensor information fusion theory in the field of rotor dynamic balancing technology. Holospectrum, which is formed by integration of the spectral lines obtained by the FFT of the rotor vibration signals collected from two mutually perpendicular radial directions, can be seen as a kind of orbit spectrum or modified FFT spectrum containing phase information. Therefore, the Holospectrum is actually an information fusion method implemented

Y. Liao (✉) · L. Qu
School of Mechanical and Engineering, Xi'an Jiaotong University,
Xi'an, People's Republic of China
e-mail: yhliao@xjtu.edu.cn

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors,
Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_3

in frequency domain [3]. Since the Holospectrum contains all the vibration amplitude, frequency and phase information of a machine set, it is more reliable and accurate in identifying the complex faults with similar vibration response spectrum structure than conventional FFT spectrum. The rotor dynamic balancing technology is then further promoted by a new method developed on the basis of this Holospectrum technique, which is called Holobalancing. By fusing the vibration data, not only the signal analysis can be extended from one-dimension to multi-dimension, it also makes the rotor balancing process more reliable and efficient. This method breaks through the limitations of some simplifying assumptions in the traditional balancing methods. The rotor spatial precession characteristic is fully considered in the analysis process of the Holobalancing method. The introduction of genetic algorithm optimization and computer simulation technique in the Holobalancing method further increased the accuracy and efficiency of this method [4].

In the Holobalancing method, rotor vibration information, as well as the system dynamic characteristic response, is expressed in the form of Three-dimensional Holospectrum (abbreviated to 3dH). The rest part of this chapter begins with the introduction of the basic theory of Holospectrum technique. Then some key conceptions and core issues of the Holobalancing method are discussed in detail. Based on that, the basic theory of the Holobalancing method and its improvement are presented. Finally, the capability and effectiveness of this method is verified by laboratory experiment and a field application case.

2 Construction of Holospectrum

2.1 Basic Condition Required

The movement of a rotor during its operation is actually a combined motion of a rotation on its own axis and a revolution round the bearing center line and is called precession in rotor dynamics. In fact, it can be seen that there are two coplanar and mutually perpendicular probes installed on each bearing section of almost every large scale rotating machinery. This arrangement makes the machine set running condition monitoring and vibration analysis more reliable. However, if the vibration signals collected from different radial directions are only analyzed independently, a complete description of the rotor spatial precession situation could not be achieved. The influences of damping effect and system stiffness anisotropy are the reasons why vibration signals collected from different radial directions are generally also not the same. It is then of great importance to find a new information fusion approach to accurately and directly describe the overall vibration situation of the whole rotor system. This is the primary motivation for the development of Holospectrum.

In essence, Holospectrum is a kind of modified FFT spectrum. Through reconstruction of the orbits of different orders of harmonic and/or sub-harmonic components in the frequency domain, Holospectrum gives a complete description of the precession situation of the whole rotor system. However, since the information fusion are all implemented in data layer, some rigorous requirements for the vibration signals participating in information fusion must be satisfied, as follows:

(1) **Consistency of probe installation**

Unlike the traditional FFT spectrum, a two-dimensional Holospectrum (2dH) should contain all the vibration information of a rotor test section. Therefore, it is required that two coplanar and mutually perpendicular probes must be installed on every rotor vibration test section and at least one keyphasor is also necessary to get the phase information. The two probes should be perpendicular to each other and both point accurately to rotor axis along radial direction. In field applications there are two most common probe installation ways–horizontal to vertical and left 45° to right 45°, as shown in Fig. 1.

One obvious advantage of the Holospectrum is that the structure and shape of 2dH are both independent of the probe mounting orientation. This is also one of the most excellent features that normal FFT spectrum doesn't possess. In fact, as long as the installation of probes follows the rules mentioned above, the Holospectra obtained will be all the same no matter which scheme is actually adopted. As shown in Fig. 2, the left side of the figure gives the probe orientation and it can be seen that the two 2dH shown in the right side are exactly the same. This feature excludes the uncertainty brought by probe orientation to vibration waveform and spectrum and makes fault diagnosis much more reliable.

As to the construction of 3dH, it is required that all the probes mounted on every test section along the shaft train should follow one same scheme. This is the consistency demanded for probe installation.

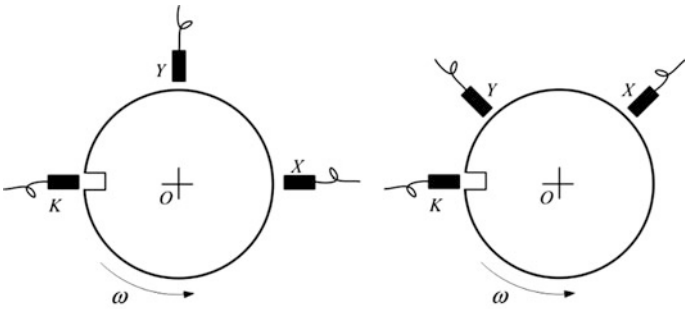


Fig. 1 Common probe installation

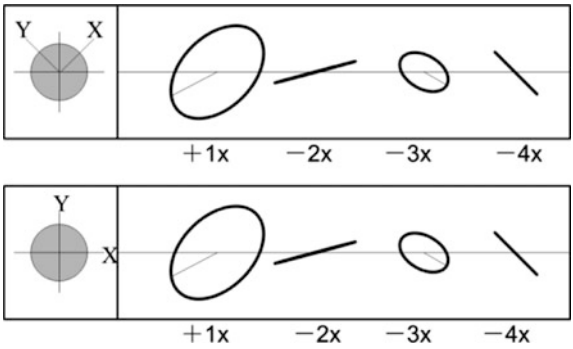


Fig. 2 2d-Holospectra of different probe orientations

(2) **Consistency of probe characteristic**

This requires that the probes used in one machine set should all have identical physical characteristic, including the linear response range, sensitivity coefficient, etc.

(3) **Consistency of signal transmission path**

This requires that the vibration signals should be transmitted via similar paths from its source to the test probes. This ensures all signals have identical transfer functions and signal characteristics during transmission.

(4) **Consistency of sampling frequency**
(5) **Consistency of the sampling start time**

This requirement ensures all signals collected have a reliable phase relationship, which corresponds to one common sampling start time. It is of vital importance to the construction of 3dH that the sampling of all signals begins at the same moment.

(6) **Accuracy of fusion information**

The information involved in the fusion process includes the amplitude, phase and frequency of all the related components of the vibration signals. Considering the fact that the characteristic frequency of some faults (such as oil whirl) may locate in the sub-harmonic area, no matter which sampling method is applied the spectral peak interpolation correction technique is always necessary to ensure the correctness of related information.

2.2 Three-Dimensional Holospectrum (3dH)

In the Holobalancing method, the rotor vibration response (including the Transfer Matrix, i.e. the rotor characteristic response) appears in the form of 3dH. 3dH is constructed by connecting all the rotational frequency orbit of the whole rotor system with generating line. The relationship between 2dH and 3dH is given in Fig. 3.

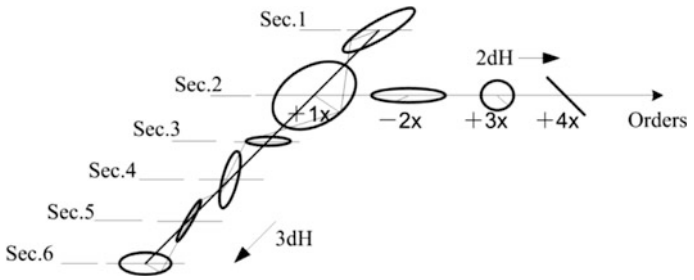


Fig. 3 Relationship between 2dH and 3dH (the example rotor contains 6 test sections)

It can be seen that only the rotational frequency components (i.e. 1X harmonic or the 1st order component) in the 2dH of every test section are used to form 3dH of the whole rotor system. This is because the 1X harmonic frequency components are the characteristic vibration response of rotor mass unbalance. For the convenience of further discussion, only these 1X components are considered hereinafter.

Suppose there are n bearing test sections along the rotor of the whole machine set and the 1X rotational frequency components, which are obtained by spectral peak correction if necessary, of the ith bearing test section are

$$\begin{cases} x_i = A_i \cos(\omega t + \alpha_i) \\ y_i = B_i \cos(\omega t + \beta_i) \end{cases} \tag{1}$$

Equation (1) can also be seen as the parametric equation of the elliptical precession orbit of the rotor axis at the ith test section, which can be further expanded to

$$\begin{cases} x_i = sx_i \sin(\omega t) + cx_i \cos(\omega t) \\ y_i = sy_i \sin(\omega t) + cy_i \cos(\omega t) \end{cases} \tag{2}$$

Equation (2) indicates that, if system rotational speed ω is constant, any elliptical orbit can be completely determined by the coefficients of the Sine term [sx_i, sy_i] and Cosine term [cx_i, cy_i].

Therefore, the 3dH of the whole rotor system can be expressed in matrix form, as shown in Eq. (3).

$$\boldsymbol{R} = \begin{bmatrix} sx_1 & cx_1 & sy_1 & cy_1 \\ sx_2 & cx_2 & sy_2 & cy_2 \\ \vdots & \vdots & \vdots & \vdots \\ sx_n & cx_n & sy_n & cy_n \end{bmatrix} \tag{3}$$

Besides, 3dH can also be expressed in a more intuitive way. Figure 4 gives the construction procedure of a 3dH of a 300 MW steam turbine generator set. Here the whole machine set contains one high pressure cylinder, one intermediate pressure cylinder, two low pressure cylinders and one generator. There are six rotors and ten bearing test sections along the shaft train. Using generating lines to orderly connect the points on every two adjacent orbits sampled at the same moment, we get the 3dH in graphical form. Figure 5 gives a typical 3dH of a rotor system with mass unbalance.

Figure 5 and Eq. (3) are two different kind of expressions of 3dH and they can be used in different occasions. Obviously, the graphical form (like Fig. 5) is more intuitive and shows the specific vibration fault type directly, while the matrix form (Eq. 3) is more convenient in balancing analysis and calculation.

For clarity only the first four sections are presented here in Fig. 5. The three essential elements necessary for the construction of graphical formed 3dH are:

(1) **Rotational Frequency Orbit**. Generally, the precession orbits are ellipses of different size and eccentricity. The size of the orbit, represented by the half length of major axis of the elliptical orbit, reflects the level of vibration and the eccentricity of the elliptical orbit gives the system stiffness anisotropic situation at different bearing test section. Besides, the inclination of the orbit major axis indicates the weak stiffness direction of corresponding bearing test sections.

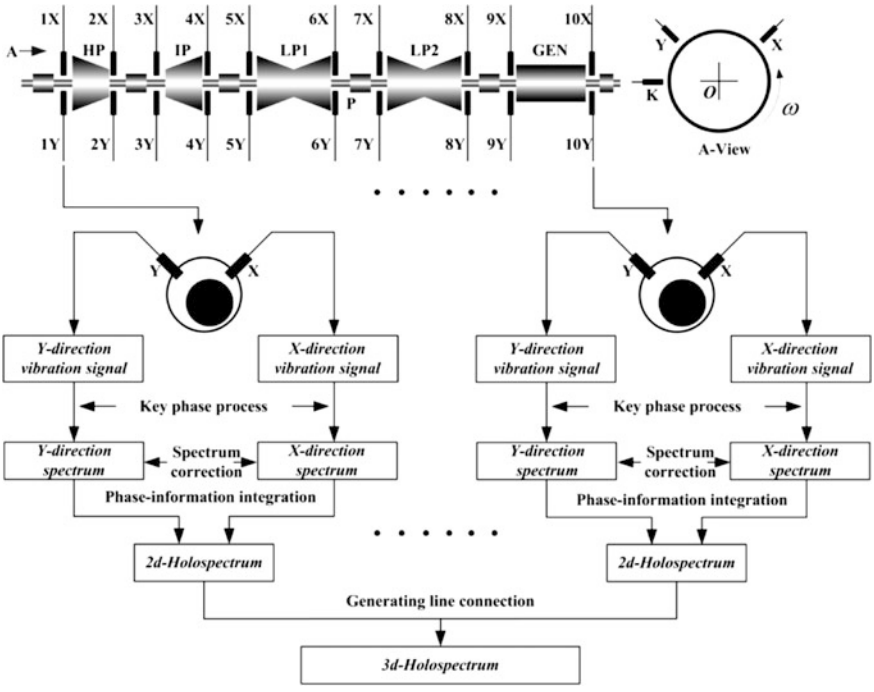


Fig. 4 Construction procedure of 3dH

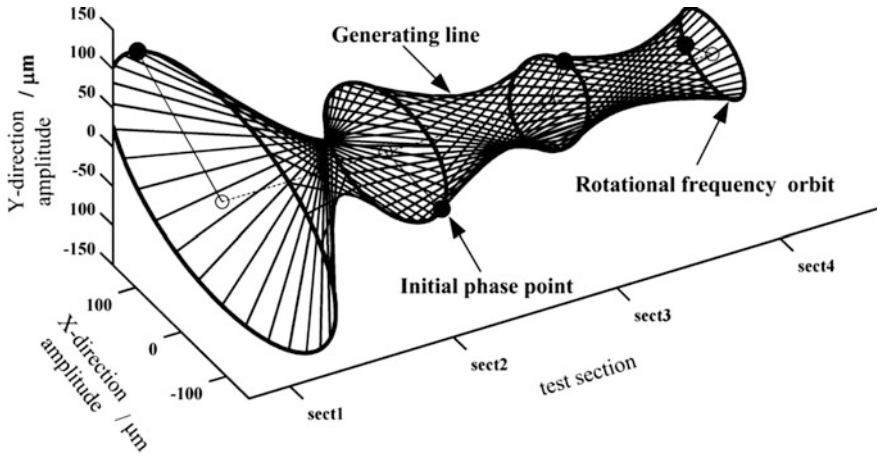


Fig. 5 The graphical form of a typical 3dH

(2) **Initial Phase Point (IPP)**. IPP gives the rotor axis location on the precession orbit at the moment when the keyphasor aims exactly at the key slot on the rotor. IPP is also the first point sampled when the data sampling process starts. Since the location change of IPP on the orbit directly reflects the change of rotor balancing state, the tight relationship between IPP and the heavy point of mass unbalance is the most important foundation for the establishment of the Holobalancing method. Besides, the vector points from orbit center to the IPP is named Initial Phase Vector (IPV). Both IPP and IPV are very important conceptions in Holobalancing method and their characteristics will be discussed in detail in the upcoming subsection.

(3) **Generating Line**. The shape of the 3dH traced out by generating lines gives us the initial perception on the way how the rotor is unbalanced. Generally, parallel generating lines may indicate the vibration problem is mainly caused by force unbalance, while crossed generating lines could suggest couple unbalance as the dominant fault. The situation of neither parallel nor crossed generating lines may give evidence of existence of multiple mass unbalance modes. In short, the shape of 3dH gives us the intuitive basis for the judgment of rotor mass out-of-balance situation.

3 Introduction of Holobalancing Method

3.1 Initial Phase Point (IPP)

Another two correlated important conceptions connected with IPP are the Initial Phase Vector (IPV) and the Initial Phase Angle (IPA, the angle of the IPV).

The magnitude of the IPV gives the measure of rotor mass unbalance and the IPA marks the spatial orientation of mass unbalance on the rotor. In order to make the relationship between IPP (including IPV and IPA) and the heavy point of rotor mass unbalance clearer, a simple rotor mass unbalance fault is considered here.

Suppose a known mass unbalance, with magnitude ε and position angle α_w, is intentionally seeded in a zero state rotor-bearing system. Then the parametric equations of its axis precession orbit, Eq. (1), can be modified to

$$\begin{cases} x = \varepsilon\lambda_x \cos(\omega t + \alpha_w + \varphi') \\ y = \varepsilon\lambda_y \cos(\omega t + \alpha_w + \psi') \end{cases} \tag{4}$$

Parameters in Eq. (4) are:

ε The amount of mass unbalance (g);
λ_x System amplification factor in x direction (μm g^{-1});
λ_y System amplification factor in y direction (μm g^{-1});
ω rotational speed (rad s^{-1});
α_w Position angle of unbalance heavy point (deg);
t Sequence of sampling time (s);
φ' Mechanical phase lag in x direction vibration (deg);
ψ' Mechanical phase lag in y direction vibration (deg).

According to the basic theory of rotor dynamic, some conclusions can be drawn. Firstly, if the dynamic characteristic of the rotor system are all constant, the eccentricity of the elliptical orbit will be unchanged during rotor operation; Secondly, if the weak stiffness direction is also constant at the same time, the inclination angle of orbit major axis will keep steady as well. It means that, as long as the working condition (including rotating speed, support stiffness, etc.) of the rotor system remain stable, related system parameters λ_x, λ_y, φ' and ψ' will all be constants and they don't vary with changes of rotor balancing state. Therefore, let $t = 0$ in Eq. (4) and we get the coordinate of IPP under this condition, as shown in Eq. (5)

$$\begin{cases} x_0 = \varepsilon\lambda_x \cos(\alpha_w + \varphi') \\ y_0 = \varepsilon\lambda_x \cos(\alpha_w + \psi') \end{cases} \tag{5}$$

Equation (4) can be further rewritten in complex form

$$\boldsymbol{r} = \mathrm{x} + jy \tag{6}$$

Therefore, the magnitude and angle of IPV, denoted by r_0 and α_0 respectively, are

$$r_0 = (x_0^2 + y_0^2)^{\frac{1}{2}}; \quad \alpha_0 = \arctan\frac{y_0}{x_0}. \tag{7}$$

Put Eq. (5) into Eq. (7), we have

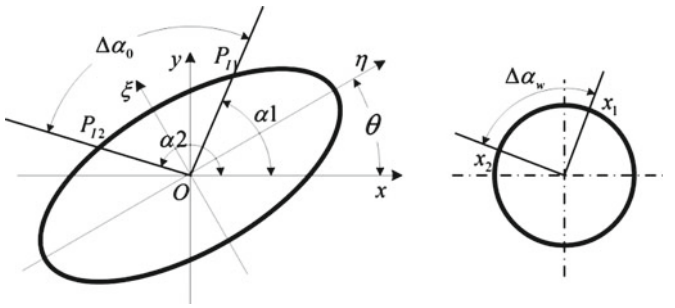
$$\begin{cases} r_0 = \varepsilon\left[\dfrac{\lambda_x^2 + \lambda_y^2 + \lambda_x^2\cos 2(\alpha_w + \varphi') + \lambda_y^2\cos 2(\alpha_w + \psi')}{2}\right]^{\frac{1}{2}} \\ \alpha_0 = \arctan\dfrac{\lambda_y\cos(\alpha_w + \psi')}{\lambda_x\cos(\alpha_w + \varphi')} \end{cases} \tag{8}$$

Obviously, r_0 and α_0 are determined by and only by ε and α_w of the unbalance heavy point. The change of rotor balancing state will definitely affect the IPP of the system response and vice versa. Therefore, this corresponding relationship between IPP and rotor mass unbalance makes it possible to use IPP (or IPV) to measure rotor system balancing state. The IPV is the key parameter in the balancing analysis and calculation of the Holobalancing method. Experiments and field applications have all confirmed this important characteristic and it will not be described here.

3.2 Precession Angle Compensation

According to the isotropic stiffness assumption of traditional dynamic balancing method, rotor axis precession orbit should be a perfect circle. That means, if the unbalance heavy point changes its position on the rotor with angle increment $\Delta\alpha_\omega$, the corresponding vibration response collected at any radial directions should all have equal amount of phase angle change, as shown in Eq. (8). In other words, if $\lambda_x = \lambda_y$ and the phase difference between φ' and ψ' is exactly 90°, which are exactly the conditions required by isotropic assumption, rotor axis precession orbit will be circular and the initial phase angle change $\Delta\alpha_0$ will be equal to $\Delta\alpha_\omega$ at any rotor test section.

Under this condition the rate of rotor rotation and that of its revolution are the same. However, this is not the case in practice. Since almost all real rotor-bearing systems have anisotropic damping and stiffness, the eccentricity ratio and major axis inclination angle of precession orbits of a rotor generally are seldom the same. Not only the phase change of any one vibration signal could not be equal to $\Delta\alpha_\omega$, but the phase change of signals collected at different axial positions are also different. Furthermore, the initial phase angle could even be changed nonlinearly, which can also be seen in Eq. (8). The change of α_0 (i.e. $\Delta\alpha_0$) is actually in a much more complex form. Therefore, angle must be compensated when we use measured $\Delta\alpha_0$ to determine $\Delta\alpha_\omega$ by taking those factors, including orbit eccentricity ratio and major axis inclination angle, into consideration.



(a) Precession angle increment $\Delta\alpha_0$ (b) Rotation angle increment $\Delta\alpha_\omega$

Fig. 6 Relationship between rotation angle and precession angle

Figure 6 gives the relationship between rotation angle and precession angle. Suppose we need to move the IPP from P_{I1} to P_{I2} according to balancing requirement. The angle compensation is implemented as follows.

The IPA of P_{I1} and P_{I2} are denoted by $\alpha 1$ and $\alpha 2$ respectively. Then the increment of IPA during this process is

$$\Delta\alpha_0 = (\alpha_2 - \theta) - (\alpha_1 - \theta) \tag{9}$$

where θ is the inclination angle of orbit major axis. The unbalance heavy point on the rotor corresponding to the IPP P_{I1} is marked by x_1.

In order to move IPP from P_{I1} to P_{I2}, i.e. IPA has an increment of $\Delta\alpha_0$ as shown in Eq. (9), the unbalance heavy point need to be moved from x_1 to x_2. The corresponding angle variation of the unbalance heavy point, $\Delta\alpha_\omega$, is

$$\Delta\alpha_w = \arctan\{(a/b)\tan(\alpha_2 - \theta)\} - \arctan\{(a/b)\tan(\alpha_1 - \theta)\} \tag{10}$$

where a and b are the length of half major and minor axis of the orbit (in μm), respectively. Equation (10), together with Eq. (9), gives the clear relationship between the change of precession angle and rotation angle, which provides the theoretical basis for angle compensation in the Holobalancing method. Since this compensation process truly considers the actual rotor precession situation under system anisotropic condition, the unreasonable isotropic assumption adopted in traditional balancing method is no longer required therefore.

3.3 *Differential Holospectrum and Transfer Matrix*

Both the Holobalancing method and the other traditional dynamic balancing methods have one common premise. That the relationship between unbalance

excitation force and system vibration response is thought to be linear is held by all dynamic balancing methods at present. More particularly, this linear premise actually contains the following two aspects of meaning. Firstly, it means that the relationship between excitation and response is linear; Secondly, it also indicates that system response satisfies linear superposition condition, i.e. system response to a group of unbalance weights is the linear combination of system responses to those unbalance weights respectively.

The correctness and reliability of this linear premise have been proved theoretically and experimentally. It sets the basic theoretical foundation for rotor dynamic balancing methods and can be utilized to extract system characteristic response, which reflects the dynamic property of the rotor system and is independent of specific out-of-balance situation. As to the Holobalancing method, the extraction of system characteristic response is very simple with only matrix addition and subtraction. Suppose the system initial unbalance response and the system trial weight response are denoted by $\boldsymbol{R}$ and $\boldsymbol{R}_t$, respectively. Then we have

$$[\boldsymbol{R}_t] = [\boldsymbol{R}] + [\Delta \boldsymbol{R}] \tag{11}$$

where $\Delta \boldsymbol{R}$ is the pure system response excited by and only by the seeded trial weight. $\Delta \boldsymbol{R}$ is also called Differential Holospectrum and can be solved directly with

$$[\Delta \boldsymbol{R}] = [\boldsymbol{R}_t] - [\boldsymbol{R}] \tag{12}$$

$\Delta \boldsymbol{R}$ depicts how the system response will be affected if the rotor balancing state varies. Another important parameter, the Transfer Matrix (TM), then could be obtained through normalization of $\Delta \boldsymbol{R}$. In the Holobalancing method, TM is defined as the system response to a standard trial weight, which has the standard mass (generally 1000 g for large scale steam turbine, for instance) and is seeded in the standard position (for the convenience of later calculation, 0° is mostly used) at the selected balancing plane. It has been proved that TM have some important features, which are also owned by the Influence Coefficient. Firstly, there is a one-to-one correspondence between TM and the balancing plane; Secondly, TM reflects the inherent dynamic characteristic of the whole rotor-bearing system. Therefore, TM is independent of specific rotor balancing state. Finally, TM is also closely related to rotating speed. TM plays a similar role of the Influence Coefficient in the dynamic balancing analysis and calculation. However, since TM contains much more information, the balancing procedure is therefore more efficient and reliable than traditional methods.

3.4 The Balancing Procedure

The problem-solving procedure of Holobalancing method is much like that of the traditional Influence Coefficient balancing method. The differences are that

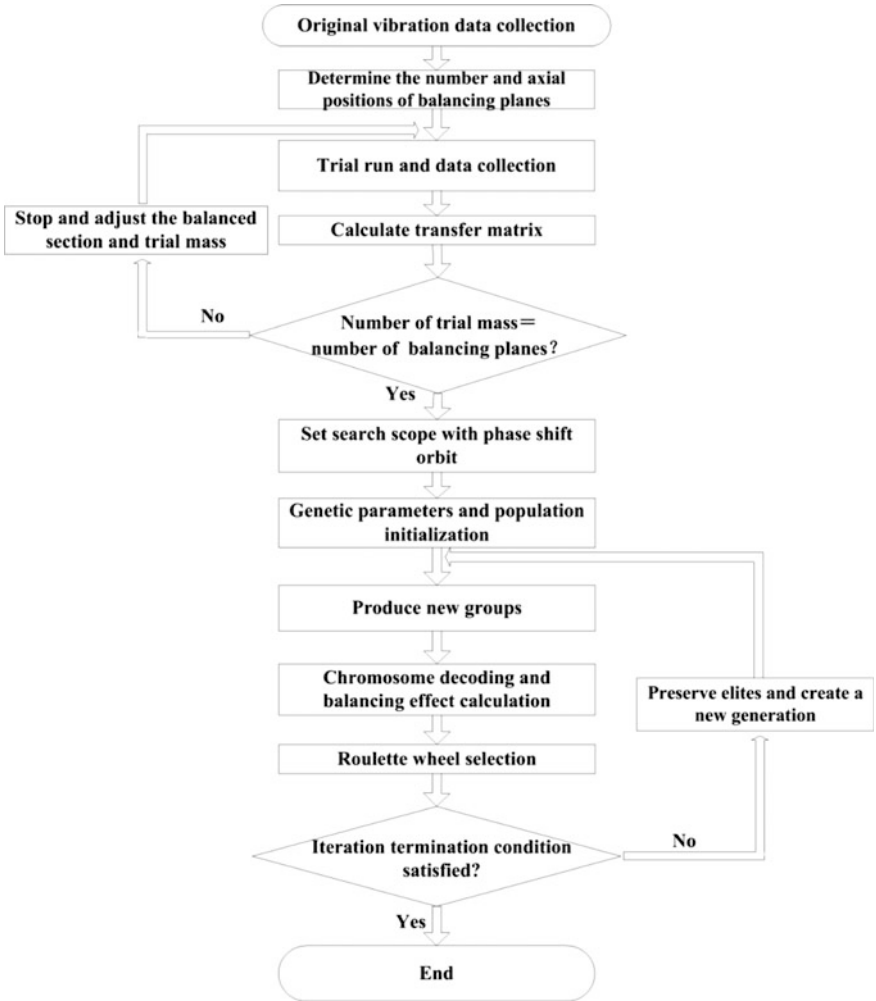


Fig. 7 The complete balancing calculation procedure of Holobalancing

vibration data of some single rotor radial direction and the Influence Coefficient are replaced by 3dH and the Transfer Matrix. Besides, since the elliptical orbits induced precession angle compensation has to be considered in the Holobalancing method, the balancing calculation procedure now is actually a nonlinear optimization problem. Many optimization techniques can be used to search for the balancing scheme, here the genetic algorithm is applied.

In summary, the complete balancing calculation procedure is shown in Fig. 7.

4 Balancing Object Reselection

Instead of the vibration vector of some single radial vibration vector, the Holobalancing method adopts the IPV as the balancing object to implement the balancing analysis and calculation, which has been proved to be effective in field balancing applications. However, since there isn't a linear relationship between the angle of the mass unbalance and the IPV, an angle compensation procedure is necessary during the calculation procedure to ensure a correct balancing scheme [5]. Furthermore, the magnitude of the IPV is not solely decided by the amount of the unbalance mass. It could also be changed by the angle variation of the heavy point. This reveals a nonlinear relationship between the mass unbalance and the rotor response (i.e. IPV) so that the computational procedure is much more complicated than the traditional methods. A more reasonable balancing object, called the Forward Precession Component of the IPV(IPV_+), is proposed here to replace the IPV.

4.1 Characteristic and Deficiency of the IPV

Suppose there are two coplanar and mutually perpendicular probes installed on each bearing section and the signals collected by them are denoted as x and y respectively. The synchronous frequency components of the vibration signals, which are the vibration occurring at 1X RPM and the main consideration in rotor balancing, collected at these two radial directions are

$$\begin{cases} x = A\cos(\omega t + \varphi) \\ y = B\cos(\omega t + \psi) \end{cases} \tag{13}$$

where ω is the rotating frequency. With Eq. (13) the synchronous shaft orbit can be constructed in a Cartesian Coordinates System as shown in Fig. 8.

Generally, the orbit is an ellipse. Since the difference between the initial phase angles of the two signals, φ and ψ, is not always exactly 90°, the coordinates system xoy constructed along with the probes will not be consistent with the one

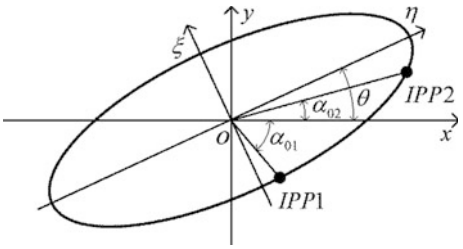


Fig. 8 The shaft orbit and the detecting coordinates system

constructed along with the major-minor axis $\eta o \xi$ of the orbit. There is an oblique angle θ between the major axis $o\eta$ and the axis ox. So neither A nor B in Eq. (13) is equal to the length of the half major axis of the orbit. It indicates that, no matter which radial direction is adopted in the balancing calculation, the result could be incorrect.

Different from those traditional balancing methods, the Holobalancing method uses the IPV of the synchronous shaft orbit as the balancing object. It has been proved that there is a certain relationship between the IPV and the mass unbalance [6], as we have discussed above in Sect. 3. The change of rotor balancing state will result in a corresponding change of the IPV. In the Holospectrum technique, the Initial Phase Point (IPP) gives the shaft center position on the synchronous shaft orbit at the moment when the key slot on the rotor aims exactly at the key phase probe during every precession cycle ($t = 0$). Then the IPV is defined as the vector starting from the orbit center and ending at the IPP. For the orbit expressed in Eq. (13), its IPP is

$$IPP(x_0, y_0) : \begin{cases} x_0 = A\cos\varphi \\ y_0 = B\cos\psi \end{cases} \tag{14}$$

Then the IPV can be expressed in complex form as

$$IPV = x_0 + jy_0 \tag{15}$$

Certainly, IPV can also be expressed in exponential form. Its amplitude r_0 and angle α_0 are shown in Eq. (7). In order to further clarify the relationship between mass unbalance and the IPV, here Eq. (8) is utilized again. It can be seen that r_0 and α_0 are dependent on ε and α_w, which suggests that the variation of IPV reflects the change of rotor balance and vice versa. Since the IPV is a better reflection of the spatial precession orbit of a rotor, it is more precise than that of some single radial direction information in dealing with rotor balancing problems.

However, since φ and ψ in Eq. (13) are also influenced by the relative installation position of the vibration probes to the key phase probe, the IPP's location on the orbit is a key phase probe orientation related. This will not be a problem if the precession orbit is an exact circle. As to an anisotropic rotor-bearing system, the precession orbit is generally an ellipse and it is possible that the IPP locates fairly close to the minor axis of the orbit while there is a large unbalance mass (see IPP1 in Fig. 8). Obviously, the IPV cannot reflect the real out-of-balance situation in that circumstance. Therefore, using IPV directly to illustrate the rotor balance could be misleading.

Furthermore, r_0 is not decided solely by ε. Suppose we keep ε unchanged but alter α_w by an increment $\Delta\alpha_w$, since the phase difference and the amplitudes of the two signals are still the same as before, the shape, size and direction of the orbit will be unchanged and only the IPP will shift along the orbit to a new position. For instance, the IPP could shift from its original position IPP1 to IPP2 after the

variation of α_w as shown in Fig. 8. The change of α_w can also result in a corresponding variation of r_0, which can be seen in Eq. (8). Though the variations of IPV can still reflect the change of the location of the heavy point, its function of illustration of unbalance response is obviously disturbed. So unbalance response should not be estimated by the magnitude of the IPV only. In addition, there is no linear relationship between the increment of α_w and α_0. According to Eq. (8), $\Delta\alpha_0 = \Delta\alpha_w$ is tenable if and only if $\lambda_x = \lambda_y$ and $\varphi' = 90° + \psi'$, which happens to be the isotropic stiffness assumption made by traditional balancing methods and is not the case for the elliptical shaped orbit. The actual relationship between α_w and α_0 is in a much more complicated form, see Eq. (8). The angle compensation procedure must be considered here to solve this nonlinear problem, which makes the balancing calculation much more complicated than the traditional methods, as we have discussed in Sect. 3.2. All these arguments indicate that the decrease or increase of the magnitude of the IPV actually doesn't correspond to the improvement or deterioration of the rotor balancing state for certain. This deficiency means that the Holobalancing method still has to rely on the vibration signals from some single radial direction for balancing analysis.

4.2 Precession Decomposition

In practice, the objective of any balancing operation should be on decreasing of the major axis of the precession orbit. However, it is not feasible to use the major axis as the balancing object directly because its direction is mainly decided by the weak stiffness direction of the rotor system and it is not sensitive to the location change of mass unbalance. For the convenience of further discussion, let's start our analysis from Eq. (4) to Eq. (6).

Substitute Eq. (4) into Eq. (6) and decompose the orbit expression according to the Euler Formula, then we have

$$r = r_+ e^{j(\Omega t + \alpha_+)} + r_- e^{-j(\Omega t - \alpha_-)} \tag{16}$$

where r_+ and r_- are the radii of the forward and backward precession circles, respectively. Equation (16) shows that the synchronous shaft orbit can be expressed as a composition of a forward precession circle with its precessing direction the same as the rotor rotating direction and a backward precession circle with its precessing direction contrary to the rotor rotating direction. If r_+ is larger than r_-, the overall precession will be forward. Otherwise, the resultant precession will be backward.

The decomposition of shaft orbit into its forward and backward components means every point on the orbit can be expressed as the superposition of two corresponding points on the forward and backward precession circle respectively. The same decomposition is applicable to IPP. IPP can be decomposed into two components, one on forward and the other on backward precession circle,

named the Forward Precession Component (IPP_+) and the Backward Precession Component (IPP_-) respectively. For any orbit, this relationship is exclusive and reversible. So the IPV can also be expressed as the vector sum of the IPV_+ and IPV_-, which can be expressed as

$$\begin{cases} IPV_+ = r_+ e^{j\alpha_+} \\ IPV_- = r_- e^{j\alpha_-} \end{cases} \tag{17}$$

The magnitude of IPV_+ and IPV_- are

$$\begin{cases} r_+ = \varepsilon \left[\dfrac{\lambda_x^2 + \lambda_y^2 + 2\lambda_x\lambda_y \sin(\varphi' - \psi')}{2} \right]^{\frac{1}{2}} \\ r_- = \varepsilon \left[\dfrac{\lambda_x^2 + \lambda_y^2 + 2\lambda_x\lambda_y \sin(\psi' - \varphi')}{2} \right]^{\frac{1}{2}} \end{cases} \tag{18}$$

The Initial Phase Angles (IPA) of IPV_+ and IPV_- are

$$\begin{cases} \alpha_+ = \arctan\left(\dfrac{\lambda_x \sin(\varphi') + \lambda_y \cos(\psi')}{\lambda_x \cos(\varphi') - \lambda_y \sin(\psi')}\right) + \alpha_w \\ \alpha_- = -\arctan\left(\dfrac{\lambda_x \sin(\varphi') - \lambda_y \cos(\psi')}{\lambda_x \cos(\varphi') + \lambda_y \sin(\psi')}\right) - \alpha_w \end{cases} \tag{19}$$

The radius ratio δ is

$$\delta = \frac{r_+}{r_-} = \left[\frac{\lambda_x^2 + \lambda_y^2 + 2\lambda_x\lambda_y \sin(\varphi' - \psi')}{\lambda_x^2 + \lambda_y^2 + 2\lambda_x\lambda_y \sin(\psi' - \varphi')} \right]^{\frac{1}{2}} \tag{20}$$

The length of the half major axis a and half minor axis b are

$$a = \left(1 + \frac{1}{\delta}\right) r_+ ; \quad b = \left(1 - \frac{1}{\delta}\right) r_+ \tag{21}$$

4.3 *Balancing Object Selection: Characteristic Analysis of IPV_+ and IPV_- [7]*

Both r_+ and r_- are unrelated to α_w and only have a linear relationship with ε (see Eq. 18). The larger the ε is, the bigger the r_+ and r_- will be, while α_+ and α_- do have a linear relationship with α_w but they are not affected by ε, see Eq. (19). α_+ and α_- will change linearly with respect to α_w. Compared with the IPV, the IPV_+ is more

direct and accurate in describing rotor balancing state. Furthermore, the radius ratio δ is independent of the mass unbalance situation and only decided by the system characteristic, see Eq. (20). That means δ is a constant during the balancing procedure and if r_+ has been effectively decreased, the actual balancing object, the major axis of the orbit, will be decreased accordingly, see Eq. (21). These features make it possible to construct a direct and reliable relationship between IPV_+ and rotor unbalance response. It is feasible to adopt IPV_+ as the correction object.

However, mass unbalance will not affect the forward precession component only. IPV_- possesses all those features as IPV_+ and could also be adopted as the balancing object according to Eq. (18) and Eq. (19). Kirk [8] once discussed this problem with Lund and they agreed that the energy generated by mass unbalance could also enter backward precession circle. It is true that the mass unbalance does affect the size of the backward precession component to a certain extent. However, it is not the decisive factor. Consider an ideal situation that the characteristic of the rotor system is isotropic. Since $\lambda_x = \lambda_y$ and $\varphi' = \psi' + 90°$, then r_- will be zero no matter how severe the out-of-balance situation is, see Eq. (18). The radius ratio δ could approach to infinity and the shape of the synchronous shaft orbit under this condition is an exact circle. It indicates that the root cause of the emergence of the backward precession component is not the mass unbalance, but the asymmetrical system characteristic. Taking this anisotropic characteristic into consideration, two different situations should be investigated when evaluating the impact of mass unbalance on the backward component.

First, suppose mass unbalance is the only fault existing in a rotor system. The radius ratio δ is a constant in such circumstances according to Eq. (20) and is determined only by the system characteristic. It means that once r_+ has been decreased (or increased), r_- will be proportionally decreased (or increased) simultaneously. Therefore, it is not necessary to bring IPV_- into consideration during the balancing procedure.

Second, consider a more common situation where there are other faults besides unbalance in a rotor system. Those faults are also featured in rotating frequency. The overall vibration response is a combination of the contributions from all these faults. An experiment was designed to help better understand the impact of the mass unbalance on the forward and backward components of the response in this circumstance. The test rig layout and the sensor locations are shown in Fig. 9.

This two mass rotor system is supported by oil-impregnated bronze bearings and connected to the motor with a flexible coupling. The rotor shaft has a nominal diameter of 10 mm, an overall length of 550 mm, and a span between bearings of 325 mm. Two disks with a diameter of 75 mm are separated within the bearing span. Each disk has 16 screw holes, equally distributed on a circle with a radius of 32 mm, for adding trial weights. The rotating direction is counterclockwise when looking at the rig from the driving end. The first order critical speed of the rotor system is around 2500 r/min and the working speed is 5000 r/min, far away from adjacent critical speeds.

There are two correction planes positioned within the bearing span, labeled as P1 and P2, respectively. The rotor system has two faults, unbalance and misalignment,

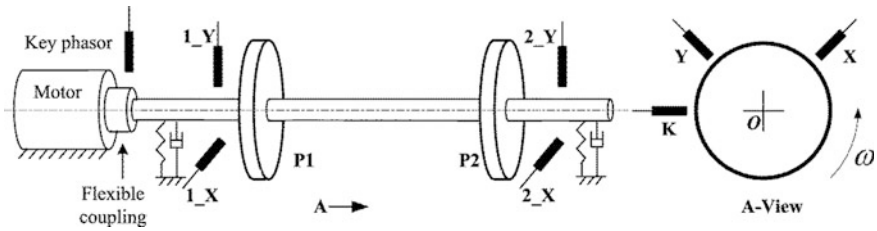


Fig. 9 Test rig layout and arrangement of the probes

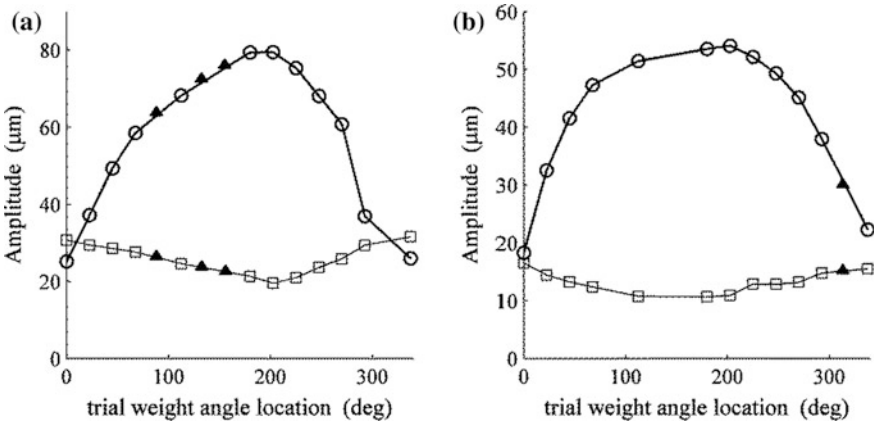


Fig. 10 The trend of the precession components (r_+: –○–; and r_-: –□–) at **a** section 1 and **b** section 2 [7]

aiming to simulate a faulty machine with multiple fault coexistence situation. In the experiment, a trial weight was sequentially added in these holes, except for those having been occupied by seeded unbalance weights. The location change of the trial weight on the correction plane corresponds to the change of the rotor unbalance state. The radius variations of the forward and backward components during this period are plotted in Fig. 10. The solid black triangles denote the positions occupied by the seeded unbalance weights. Along with the movement of the trial weight on the correction plane, r_+ has a variation of nearly 70 μm in Sec. 1 and 50 μm in Sec. 2, while during the same period the variation in r_- at both sections is insignificant and even shows a slightly opposite trend with that in r_+ in this case.

Obviously, the radius ratio δ is not a constant anymore. The change of δ during the balancing procedure indicates there must be some faults other than mass unbalance. Equations (18)–(21) are not valid under this condition. To explain this phenomenon, the effect of misalignment should be considered. Actually, all faults featured in 1X frequency can excite both forward and backward components when the system characteristic is asymmetrical [9], not just misalignment and unbalance mentioned in this example. The amount of IPV_- depends not only on the

characteristics of the rotor-bearing system but also on those of the specific faults. The emergence of large eccentricity orbit, or large backward component, does not suggest severe unbalance for sure. Since the overall IPV_+ and IPV_- are combinations of these two faults, they can be expressed in a separated form as

$$IPV_+ = IPV_{u+} + IPV_{n+}; \quad IPV_- = IPV_{u-} + IPV_{n-} \tag{22}$$

where IPV_{u+} and IPV_{u-} are the components induced by mass unbalance; IPV_{n+} and IPV_{n-} are the components generated by misalignment. It will be difficult to actually extract IPV_{u+} and IPV_{u-} from IPV_+ and IPV_-; thus only a qualitative analysis is carried out here.

Suppose IPV_{u+} and IPV_{u-} are decreased to IPV'_{u+} and IPV'_{u-}, respectively after an effective balancing operation, and at the same time IPV_+ and IPV_- are changed to IPV'_+ and IPV'_-, correspondingly. The impact of the improvement of rotor balance on the IPV_+ and IPV_- then could be different, which can be explained as follows.

Since mass unbalance has the largest radius ratio among all faults featured in working frequency, which indicates the IPV_{u-} cannot be the main part of the IPV_-, the decrease in IPV_{u-} then may have little impact on IPV_-, or even result in a slight increase in IPV_- instead of decreasing it, as shown in Figs. 10 and 11.

Whether IPV_- will be decreased or increased depends on the phase relationship between the IPV_{u-} and IPV_{n-}. Though there is IPV_{n+} in the IPV_+, the impact of IPV_{n+} on IPV_+ will not be so obvious compared with the impact of IPV_{n-} on IPV_- when the primary fault has been identified as mass unbalance. Since the backward component is not sensitive to the change of the rotor balance in the compound fault situation, the decrease in IPV_{u+}, caused by the balance improvement leads to an immediate decrease in IPV_+, while the change in IPV_- is in significant.

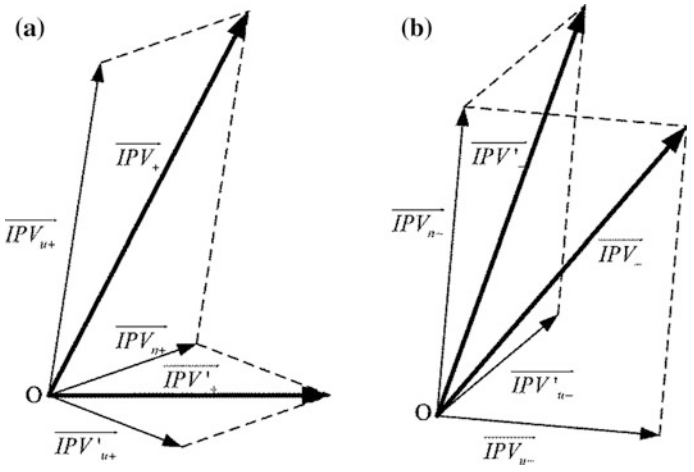


Fig. 11 The impact of the improvement of rotor balance on **a** IPV_+ and **b** IPV_- [7]

In conclusion, no matter in what circumstances, the IPV_- is not important in the balancing procedure. Furthermore, the use of IPV_+ as the balancing object has another additional benefit. Since the forward precession component is an exact circle, the isotropic characteristic is no longer an unreasonable assumption. So, the IPV_+ can also be used in the other balancing methods. This removes the obstacles for the future integration of the Holobalancing method and the other balancing methods.

5 Experimental Verification and Case Study

5.1 Experimental Verification

The test rig shown in Fig. 9 is used to validate the proposed method. Raw vibration signals collected from both bearings exhibit overwhelming 1X frequency component and the purified synchronous shaft orbits in the 3dH of system initial vibration response are both ellipses with large eccentricity, as shown in Fig. 12a. This kind of orbit shape generally indicates that there could be some faults other than mass unbalance that also need to be corrected, such as insufficient supporting stiffness. Since the phase and amplitude are stable, a balancing experiment is conducted to

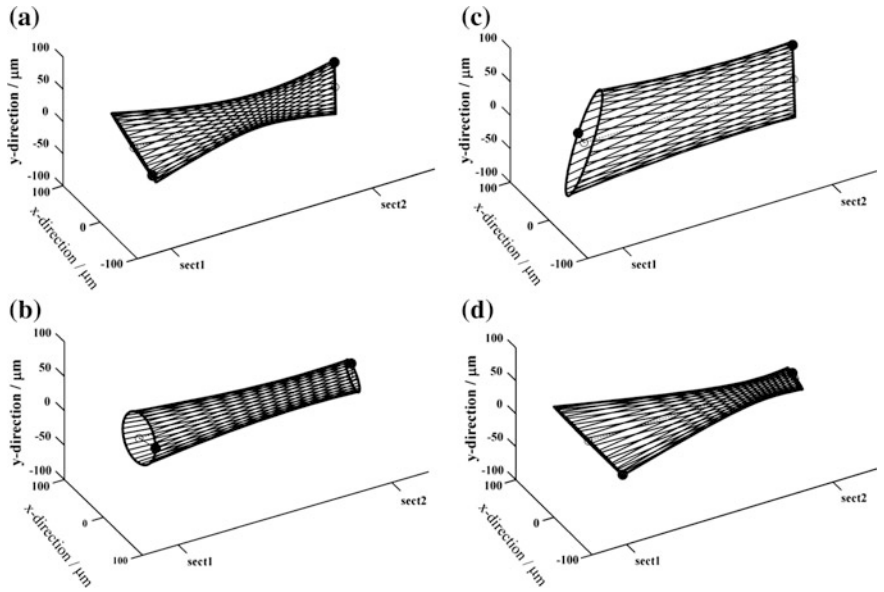


Fig. 12 **a** The system initial vibration response in 3dH [7]. **b** The balancing effect using IPV_+ scheme [7]. **c** The balancing effect obtained with information from only *x* direction [7]. **d** The balancing effect obtained with information from only *y* direction [7]

Table 1 Vibration data of the initial and trial run (μm ∠ °)

	1_X	1_Y	2_X	2_Y
Initial vibration	60.81 ∠ 144.80	20.05 ∠ 137.11	1.81 ∠ 20.20	42.18 ∠ 26.93
Trial run with 1.30 g ∠ 135° on p1	50.57 ∠ 144.42	32.06 ∠ 99.07	46.24 ∠ 44.74	43.09 ∠ 130.08
Trial run with 1.30 g ∠ 135° on p2	59.21 ∠ 96.00	60.13 ∠ 136.57	7.97 ∠ 24.27	44.27 ∠ 50.24

Table 2 Comparison of balancing solutions

Balancing method	Balancing object	Balancing solution (g ∠ °)	
		P1	P2
Holobalancing method	IPV_+	0.45 g ∠ 81.56°	0.88 g ∠ 235.63°
Influence coefficient method	*x* direction info	0.25 g ∠ 30.48°	1.60 g ∠ 200.00°
Influence coefficient method	*y* direction info	0.80 g ∠ 113.70°	0.95 g ∠ 334.50°

see whether balancing operation could effectively reduce the vibration level under this condition. Table 1 lists the synchronous vibration data of the initial and the trial runs.

The balancing solutions calculated by the traditional influence coefficient method and the Holobalancing method, which use single radial direction information and the IPV_+ respectively as the balancing object, are listed in Table 2. There are remarkable differences between the correction schemes calculated with information of different radial directions, especially the angle of the correction weights. If this kind of situation happens in real dynamic balancing issue, it will be very hard for the field engineers to decide which scheme should be chosen. That is not only because the operation cost is involved, but in some serious situation a wrong scheme could even endanger the safety of the machine and the operators.

The three balancing solutions are applied to the rotor, respectively, and their balancing effects are shown in Fig. 12b, d. The vibration in the *x* direction was visibly reduced with single *x* scheme (from 61 μm to 47 μm at Sec. 1 in Fig. 12c), but vibration in the *y* direction was greatly increased instead (from 20 μm to over 100 μm also at Sec. 1 in Fig. 12c). Similar results can also be seen in the balancing effect of single *y* scheme, as shown in Fig. 12d. That is why it is necessary for traditional balancing method to verify the balancing solution with the vibration information from other radial directions.

Moreover, although the vibration level detected at one radial direction after applying the single *x* or *y* schemes may be extremely small (even less than 10 μm), the actual vibration level is not that small as detected. In fact, it can be seen that the balancing effects, represented by the length of the major axes of the precession orbits, after the correction with the information of single radial direction (bold numbers in the last two rows of Table 3) are not decreased as expected. On the countrary, the balancing effect of the holobalancing method, as shown in the second data row in Table 3, greatly improves the vibration situation of the testrig.

Table 3 Parameter comparison

	r_+ (μm)		r_- (μm)		a (μm)	
	Sec 1	Sec 2	Sec 1	Sec 2	Sec 1	Sec 2
Initial vibration	30.71	21.22	33.26	21.00	63.98	42.24
IPV_+ scheme	5.29	5.16	39.60	16.56	44.89	21.73
x scheme	64.45	27.48	48.16	28.03	**112.58**	**55.50**
y scheme	44.89	9.01	46.87	12.17	**91.80**	21.16

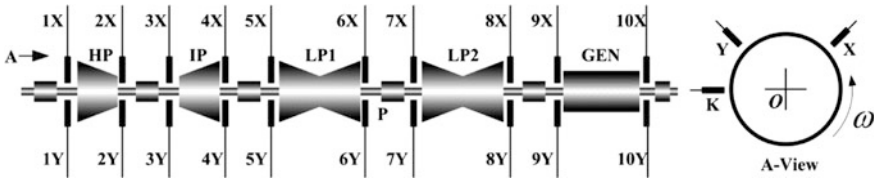


Fig. 13 The sketch map of the 300 MW turbine ge set [7]

In fact, this kind of situation is caused by the large eccentricity of the precession orbits and the sensor orientation. It is the major axis of the precession orbit that truly represents the actual vibration level. Large eccentricity orbit has a significant impact on the unbalance analysis procedure. Orbits with large eccentricity indicate that the characteristics of the rotor-bearing system in different radial directions are highly asymmetrical. Therefore, the isotropic assumption made by traditional balancing methods will not be valid. The analysis and calculation based on this assumption could be inaccurate. The Holobalancing method with the IPV_+ as the balancing object extracts the rotor unbalance information more precisely than traditional methods. With the adoption of the IPV_+, the interference caused by backward component has been eliminated. In this experiment, among all the three balancing schemes, only the scheme of the Holobalancing method effectively corrects the rotor balance. The half major axes of the precession orbits from both bearing sections have been decreased by 30 and 49%, respectively.

5.2 Case Study

Figure 13 shows the sketch map of a 300 MW turbine generator set. The overhaul of this turbine generator set had lasted nearly 2 months to replace all the broken blades. In the first startup after the overhaul, vibration from bearing sections 4–6 all exceeded the alarm level. The largest vibration peak-to-peak value was almost 200 μm.

The vibration signals were carefully analyzed and it was confirmed that the vibration was excited mainly by mass unbalance. The first trial weight (1597 g) was applied on the correction plane adjacent to sect. 5. The effect of the trial weight was

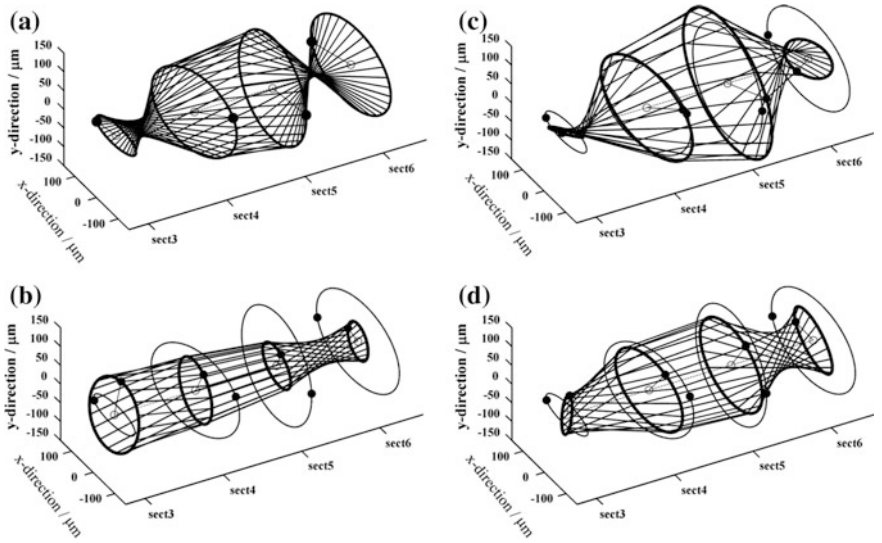


Fig. 14 **a** The initial vibration [7]. **b** The comparison between the initial vibration (*thin line*) and the final balancing result (*bold line*) [7]. **c** The comparison between the initial vibration (*thin line*) and that of the second trial run (*bold line*) [7]. **d** The comparison between the initial vibration (*thin line*) and that of the third trial run (*bold line*) [7]

Table 4 The balancing solution comparison

Balancing object	LSO	WLSO
x direction info	2109.58 g ∠ 334.64°	2123.75 g ∠ 338.05°
y direction info	1195.12 g ∠ 3.32°	1566.11 g ∠ 1.16°

negative and the vibration was increased. The machine set was even unable to pass through its first critical speed.

The initial vibration is shown in Fig. 14a (only sections 3–6 are illustrated here for convenience). After removing the first trial weight, another two weights (1384 and 1775 g) were added successively on the correction plane located adjacent to sect. 6 (labeled as the **P** in Fig. 13). The effects were still unsatisfactory. The comparisons between the initial vibration and these two trials are shown in Fig. 14c, d, respectively.

It seems that the balancing attempt to correct this rotor has reached a deadlock. Table 4 lists the balancing schemes calculated from the data collected from the third trial run using the influence coefficient balancing method with the least-squares optimization (LSO) and the weighted least-squares optimization (WLSO), respectively.

The balancing weight calculated with the information of *x* direction was much heavier than that of the *y* direction. Even in the balancing schemes with WLSO method, the difference of these two solutions is still more than 500 g. Such a big

Table 5 Comparison of balancing results

Stage of balancing procedure	Major axis (μm)			
	Sec. 3	Sec. 4	Sec. 5	Sec. 6
Original vibration	92.27	203.47	155.64	185.80
The second trial	85.45	206.30	199.40	115.87
The third trial	57.48	148.05	163.46	87.32
Final balancing	104.72	76.11	62.80	45.42

difference makes it very difficult for the field engineers to decide. The balancing solutions calculated by the Holobalancing method with IPV and IPV_+ being the balancing objects are 1822.96 g ∠ 331.98° and 1879.36 g ∠ 339.54°, respectively. Due to the limitation from the correction plane, a weight of 1835 g was finally seeded on the plane at 342°. The balancing result is depicted in Fig. 14b. Table 5 presents the unbalance response of all the four sections during the whole balancing procedure. Using the improved Holobalancing method proposed, the vibration level from sections 4–6 have been reduced greatly.

6 Conclusion and Discussion

In this field application, the calculated balancing solutions using IPV_+ and IPV are similar. One advantage of using IPV_+ as the balancing object is that since the relationship between IPV_+ and the rotor balancing state is linear, neither angle compensation nor phase shift orbit is required anymore. Therefore, the whole calculation procedure is greatly simplified without sacrificing the balancing precision. Laboratory experiment and field application have successfully demonstrated the feasibility and effectiveness of the IPV_+ based approach. At the same time, since the decreasing of the IPV_+ can decrease the major axis at the same time, there is no need to verify the correction scheme with vibration signals from different radial directions and the balancing efficiency is then improved. However, some limitations that need further investigation should be mentioned here. In fact, all faults featured in rotating frequency will excite both forward and backward precession components at the same time. The result of this study does not imply that one can discriminate different faults through this decomposition procedure, and that is also beyond the scope of this chapter. Besides, there will be two different critical speeds in the same order when the rotor system is anisotropic. The precession excited by mere mass unbalance could be obviously backward when the rotating speed is located in the region between these two different critical speeds [10]. The use of IPV_+ as the balancing object in that speed region could get into trouble. This will not be a problem if the rotor under consideration is a rigid one. As to the balancing of flexible rotors with the modal balancing method, since the mode separation has to

be done at a speed fairly close to a corresponding critical speed region, this situation then has to be carefully considered. The ongoing study is to investigate the rotor characteristic with anisotropic stiffness in this special speed region and integrate the use of IPV_+ with the modal balancing method.

References

1. Qu L., Liu X., Peyronne G, Chen Y., "The Holospectrum- A new method for rotor surveillance and diagnosis," Mechanical Systems and Signal Processing, 1989, 3:255–267.
2. Qu L., Xu G., "One decade of Holospectral technique: review and prospect," Proceedings of the 1999 ASME, Design Engineering Technique Conferences, 1999.
3. Qu L., Holospectrum and Holobalancing Technique in Machinery Diagnosis, Science Publication House, 2007.
4. Liu S., Qu L., "A new field balancing method of rotor systems based on holospectrum and genetic algorithm," Applied Soft Computing, 2008, 8:446–455.
5. Liao Y., Zhang P., Lang G., "Phase shift orbit and its application in Holobalancing procedure," Journal of Xi'an Jiaotong University, 2008, 42:803–806.
6. Qu L., Qiu H., Xu G., "Rotor balancing based on Holospectrum analysis: principle and practice," China Mechanical Engineering, 1998, 9:60–63.
7. Liao Y., Lang G., Wu F., Qu L., "An improvement to Holospectrum based field balancing method by reselection of balancing object," Journal of Vibration and Acoustics, 2009, 131: 031005–1.
8. Kirk R.G., "Lund's elliptic orbit forced response analysis: the keystone of modern rotating machinery analysis," Journal of Vibration and Acoustics, 2003, 125:455–461.
9. Liao M., "The processing analytical method for rotor vibration and its application," China Plant Engineering, 2003, 10:51–52.
10. Zhou R., Rotor Dynamic Balancing – Theory, Method and Application, Chemical Industry Publishing House, Beijing, P. R. China, 1992.

Wavelet Transform Based on Inner Product for Fault Diagnosis of Rotating Machinery

Shuilong He, Yikun Liu, Jinglong Chen and Yanyang Zi

Abstract As a significant role in industrial equipment, rotating machinery fault diagnosis (RMFD) always draws lots of attention for guaranteeing product quality and improving economic benefit. But non-stationary vibration signal with a large amount of noise on abnormal condition of weak fault or compound fault in many cases would lead to this task challenging. As one of the most powerful non-stationary signal processing techniques, wavelet transform (WT) has been extensively studied and widely applied in RMFD. Many previous publications admit that WT can be realized by means of inner product principle of signal and wavelet base. This paper verifies the essence on inner product operation of WT by simulation experiments. Then the newer development of WT based on inner product is introduced. The construction and applications of major developments on adaptive multiwavelet in RMFD are presented. Finally, super wavelet transform as an important prospect of WT based on inner product are presented and discussed.

1 Introduction

Rotating machinery is widely used in industrial equipment. For rotating machinery, some pivotal components could not avoid generating multifarious faults after running in the complex and severe conditions for a long time such as strong impact, corrosive environment, high temperature or heavy load [1]. In rotating machinery, some arisen fault may lead to catastrophic accidents as well as enormous economic losses. Therefore, it is necessary and crucial to identify the type of fault and evaluate the level of fault as early as possible, particularly on two vital problems in the rotating machinery fault diagnosis (RMFD), compound fault diagnosis and weak fault diagnosis [2]. Vibration signal analysis remains the most popular and useful method in the assignment of RMFD [3, 4]. However, when the equipment is operated, numerous kinds of mechanical faults will produce specific dynamic

S. He (✉) · Y. Liu · J. Chen · Y. Zi
Xi'an Jiaotong University, Xi'an, People's Republic of China
e-mail: xiaofeilonghe@163.com

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors, Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_4

response signals that is of diversification. Moreover, because the structures is correlative and the equipment is complex, the measured vibration signal that acquired is not only complicated but also non-stationary, and heavy background noise bury the fault features. Consequently, great difficulty exists in identifying this fault feature from such vibration signals that is acquired. Engineering requirements have promoted continuous advancement of some signal-processing technologies like short-time Fourier transform, the fast Fourier transform, wavelet transform and etc., and a foundation is established for the fault diagnosis of rotating machinery, the condition monitoring and got significant influence in this field [5].

We can treat Fourier transform (FT) [6] as well as short-time Fourier transform (STFT) [7] as kinds of inner product transform, which analyses the contents that in the signal by using a pre-determined triangular basis. Fourier transform became a signal processing method which is the most widely used as the bridge from the analysis of time domain to frequency domain [14]. But local feature information that in frequency domain as well as its corresponding relations that in the time domain can't be provided by Fourier transform [8]. To solve the problem, a more effective frequency and time localized analysis method named the STFT was proposed Gabor. And we could consider the STFT as a local spectrum of the signal in a fixed window [9]. Although STFT has achieved tremendous achievements in mechanical fault diagnosis, the effectiveness of it is still restricted by the inevitable limitation from single triangular basis, and it indicates that Fourier Transform could be good at detecting harmonic feature rather than the usual impulse feature of the fault of rotating machinery that based on the principle of inner product.

Wavelet transform (WT), unlike FT, has more basis functions to choose to match a special fault symptom, which can benefit fault feature extraction [10, 11]. The idea of translation and dilation originates the wavelet theory that is also a kind of inner product transform, which analyzes the non-stationary contents that in the signal by using a pre-determined wavelet basis [12]. On the basis of the study from Randall, it is the similarity that between the non-stationary contents that in the signals and the wavelet basis plays a conclusive role in its successfulness [13]. What's more, because of the advantage exist in multi-resolution analysis, tremendous usefulness has already been shown in fault diagnosis of rotating machinery by WT [14]. We usually categorize the wavelet transform as wavelet packet transform (WPT), discrete wavelet transform (DWT), and continuous wavelet transform (CWT) [15]. However, all of these wavelet transform methods have trouble in selecting appropriate wavelet basis. And without an inappropriate wavelet basis employing in an application, the accuracy of the fault diagnosis, particularly for weak fault diagnosis will be directly influenced.

Sweldens proposed Lifting scheme (LS) originally, which reveals itself as a systematic and flexible way that can construct second generation wavelet basis and also is dyadic wavelet basis [16]. A biorthogonal wavelet basis, a wavelet, in which the relevant wavelet transform is invertible but not always orthogonal. Compared with orthogonal wavelets, designing biorthogonal wavelets provides more degrees of freedom. And one additional degree freedom gives a possibility to structure symmetric wavelet functions [16]. Because the structure is generated from the time

domain without relying on the FT. LS, as a tool to design wavelet, provides much more freedom and flexibility for the structure of biorthogonal wavelet [17]. What's more, we can use it to construct adaptive wavelet by the design to predict operator and update operator [17]. In the light of the prediction operator, the prediction step provides the detail signal that shows the difference between prediction outcome and original signal. And an approximated signal that produced by update operators offers a general expression of original signal [17]. Lifting scheme also has several other advantages than classical wavelet transform, e.g. simple, less memory and computation, integers-to-integers wavelet transforms and irregular samples [18]. Although the DWT has achieved tremendous in mechanical fault diagnosis, the effectiveness of it is still hampered by the inevitable limitations [18]. Single wavelet basis selected during SGWT or DWT cannot extract the feature with multiple types of shapes, which become a reason for the failure of compound faults diagnosis of rotating machinery. But as the matter of fact, because of the complexity of equipment and the correlation of structures, engineering practice powerfully indicates that faults appeared in rotating machinery usually are expressed as the compound fault.

In addition to reducing frequency aliasing and improving shift insensitivity, Kingsbury propose dual tree complex wavelet transform (DTCWT) by using two different wavelet bases and it could extract two relevant features at the same time [19]. Moreover, Donovan et al. firstly propose multiwavelet transform (MT), which is a newer development of the wavelet transform theory [20]. Multiwavelet can realize multi-resolution analysis and simultaneously possesses some important properties, such as symmetry, orthogonality, higher order and compact support of vanishing moments, which traditional scalar wavelet doesn't have [21]. At first, multiple wavelet basis, depended on inner product principle, has more obvious advantages compared with traditional wavelet transform during multiwavelet transform on extracting some compound features with multiple types of shapes in rotating machinery fault diagnosis [22]. The theory of MT and DTCWT provides a possibility to solve the detection of compound faults of rotating machinery.

The world has seen an enormous growth in wavelets' theory and application, and many publications have appeared to describe wavelet theory's advancement and it successful used in a multifarious of fields of engineering [23]. These methods became a powerful mathematical and a signal processing tool to identify machine conditions in operation because of the adaptive, multi-resolution ability. But actually, in view of multiplicity and adaptivity of wavelet basis that depend on the nature of inner product principle, some recent developments like adaptive SGWT, MT and DTCWT remarkably enhance the capacity of WT on compound fault and weak fault diagnosis in rotating machinery. For better solving these two vital problems of compound fault diagnosis and weak fault diagnosis in RMFD, revealing the nature of inner product operation and after that plan the development in the future is necessary. Then we introduce the construction and the applications of main developments on accommodative multiwavelet in RMFD. Finally, we present and discuss super wavelet transform, an important prospect, which of WT that based on inner product.

2 Wavelet Transform Based on Inner Product

2.1 Inner Product

The inner product theory plays a vital part in signal processing [24]. Follows are the definition of the inner product of the function which is in the square integrable space of real numbers $L^2(R)$:

$$\langle x(t), y(t) \rangle = \int_{-\infty}^{+\infty} x(t) y^*(t) dt \tag{1}$$

The superscript * symbolizes the conjugate transposition. $\langle x(t), y(t) \rangle$ stands for an operation aimed at calculating a generalized inner product betwixt $x(t)$ and $y(t)$.

The inner product theory used in mechanical fault diagnosis is brought in from the general expansion of signals. For the signal expression is diverse, a given signal is able to be expanded in different ways. If a signal x in space of Ψ can be evinced as follows

$$x = \sum_{n \in Z} a_n \psi_n \tag{2}$$

$\{\psi_n\}_{n \in Z}$ is a cluster of fundamental functions in space of Ψ.

If $\{\psi_n\}_{n \in Z}$ is the flawless sequences in Ψ space, which means every signal in space of Ψ can be evinced as Eq. (2), there exists a cluster of dual functions, whose expansion coefficient can be obtained by basis function as in Eq. (3) or Eq. (4):

$$a_n = \int x(t) \psi_n^*(t) dt = \left\langle x, \tilde{\psi}_n \right\rangle \tag{3}$$

$$a_n = \sum_{t \in Z} x(t) \psi_n^*(t) dt = \left\langle x, \tilde{\psi}_n \right\rangle \tag{4}$$

Note that $\tilde{\psi}_n$ stands for the analytic function and ψ_n stands for the synthesis function, while each functions is dual. $\left\langle x, \tilde{\psi}_n \right\rangle$ stands for an operation of computing a generalized inner product betwixt x and $\tilde{\psi}_n$.

According to Eq. (2), a bigger result of a_n stands for a closer relationship betwixt the signal x and the dual functions $\tilde{\psi}_n$. The inner product transform can be such an easy and fixed process if the dual functions $\tilde{\psi}_n$ is considered as a kind of basis function and the inner product transform is considered as a kind of measurement of the similarity between the signal x and the basis function $\tilde{\psi}_n$. As the intrinsic quality of the mechanical fault diagnosis is to find out the mode which has the most

similarities to the basis function, the vital step of the fault diagnosis is to structure the appropriate basis function and distinct signal features [25]. As a consequence, the core of the accurate condition monitoring and fault diagnosis is extracting the realistic and physical features from the original signal through inner product transform.

FT is the most widely applied method in signal processing. It could be solved by means of the inner product operation $\langle x(t), e^{j2\pi ft} \rangle$ betwixt a signal $x(t)$ and a triangle basis function $e^{j2\pi ft}$ [25].

2.2 CWT, DWT and WPT

WT can be regarded as a fast-evolving mathematical and signal disposing tool that transforms a signal in time domain through a wavelet basis function into a diverse form, i.e. mostly a sequence of wavelet coefficients in time-scale domain [26]. The wavelet basis function is a necessity to implement the wavelet transform. And wavelet basis function is a small wave, which owns oscillating waveform like features and focuses its whole energy in short time as well. In a word, the traditional WT can be sorted as CWT, DWT, and WPT.

Let us first state some annotations and summarize the very rudiments on wavelet transform theory. In the meantime, the extensive literature is capable to provide further particulars on the wavelet theory, refer to example [27]. With the parameter $s > 0$, a general wavelet dictionary $\{\psi_{u,s}\}$ can be defined as the dilated in the temporal domain and interpreted by $u \in R$ of the mother wavelet ψ (of zero-mean) as follows

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) \tag{5}$$

Then the wavelet transform of $f = x(t)$ is realized by computing the inner product operation of $W_f(u,s) = \langle f, \psi_{u,s} \rangle$.

$$W_f(u,s) = \langle f, \psi_{u,s} \rangle = \frac{1}{\sqrt{s}} \int x(t) \psi\left(\frac{t-u}{s}\right) dt \tag{6}$$

If s stands for a continuous variable, in that way $W_f(u,s)$ is defined the CWT but if $s = a^j$, a is the scale parameter, in that way $W_f(u,s) = W_f(u,j)$ is defined the DWT [27]. Discrete wavelet transform can be treated as the application of a filter bank is an important property of it, which means that each filter corresponds to one scale. In practice, the most applied case is prescribed by dyadic subdivision scheme, $s = 2^j$ [28].

In 1984, wavelet as a novel notion was first definitely submitted by Morlet. However, this notion still faces a lot of dispute and criticism for the time. Soon

after, Morle formalized the plan of CWT and inferred its inverse transform with the aid of Grossmann [29]. In 1985, a flawless orthogonal wavelet basis was established by Meyer and this wavelet basis processes quite excellent time and frequency localization property, and it was significant to extract local essential information of a signal and developed the engineering applications of CWT [30]. In the following year, Meyer and Mallat proposed the thought of multi-resolution analysis (MRA) which urged that to construct more orthogonal wavelet basis is very convenient and easy [31]. A more significant affair was that the MRA had given birth to the well-known fast DWT, which was used to computing the wavelet transform coefficients of the signal which is based on recursive filtering [32]. The flow chart of three-stage DWT decomposition and reconstruction processes is expressed in Fig. 1. Before long, Daubechies built orthogonal wavelet basis with contract support property in an original way [33]. Moreover, Daubechies also did lots of researches on structuring wavelet frames which allowed more freedom about the design of basis wavelet function with a little cost of some redundancy [34]. These researches from Mallat and Daubechies vastly promoted the progress of the wavelet theory from continuous to discrete signal analysis and engineering applications.

In 1992, Coifman, Meyer and Wickerhauser suggested the algorithm framework of WPT, which was regarded as a natural and significative expansion of the MRA and DWT [35, 36]. As we are always concentrating on developing comprehensive representations of a signal, probably the most known way is the WPT based on a basis pursuit framework through successive scale refinements of the expansion [35, 36]. WPT can further decompose the particular coefficients of the analysed signal in the high frequency part and offers a more particular and comprehensive time-frequency plane tiling as it is shown in Fig. 2.

WT, unlike FT and STFT, has more basis functions to choose to match a special fault symptom, which can benefit fault feature extraction. What's more, because WT, which is different from them, can be used to multi-scale analysis of the signal via dilation and translation, it can effectively identify the time–frequency feature of

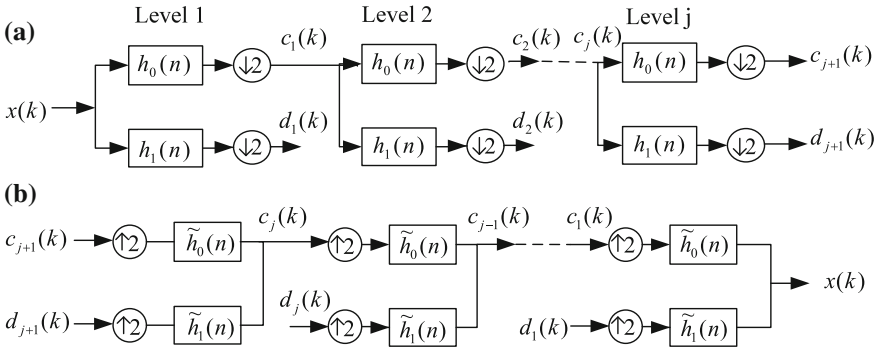


Fig. 1 The flow chart of **a** three-stage DWT decomposition and **b** reconstruction

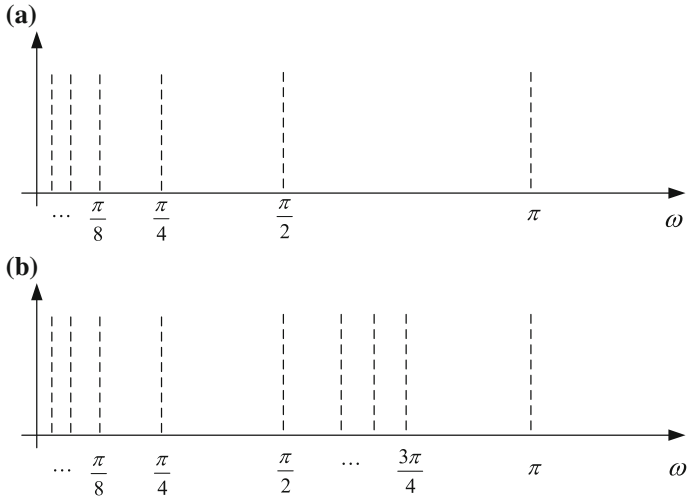


Fig. 2 **a** Dyadic WT time-frequency plane tiling and **b** dyadic WPT time-frequency plane tiling

the signal. Consequently, WT has obvious advantages when analyze non-stationary signals. Recently, WT have obtained great success in the fault diagnosis of rotating machinery because of its distinct advantages, not only for the ability of the analysis of the non-stationary signals. During the past ten years, People have made great progress in the theory and the applications of wavelet theory as well as in the field of RMFD. Nevertheless, for better solving these two vital problems of compound fault diagnosis and weak fault diagnosis in RMFD, revealing the nature of inner product operation and after that plan the development in the future is necessary.

2.3 Inner Product Validation of WT in RMFD

He et al. design and use the Daubechies 10 (Db10) wavelet to emulate the weak characteristic of the influence fault in the mixed analog signals [24, 33] to justify that the basis function is most similar to the feature can perfectly match the mechanical fault characteristic. In Fig. 3, its abscissa represents the quantity of samples N and its ordinate reflects the non-dimensional amplitude A, and the Db10 wavelet function is shown in it. The simulated signal is generated by adding a small impulse component, the Db10 wavelet function $\psi(t)$, to the sinusoidal signal which imitates the operating feature of vibration signal on rotating machinery as shown. The sample frequency of the analog signals is 5120 Hz and the quantity of sampling points is 5120, which meets the sampling needs of the impulse component and the sinusoidal component. The Fig. 4 reflects the periodic impulse component $x_1(t)$, the sinusoidal component $x_2(t)$ and the mixed simulated signal $x(t)$. In the Fig. 4a, c, the first simulative impulse in the Db10 wavelet function begins from the

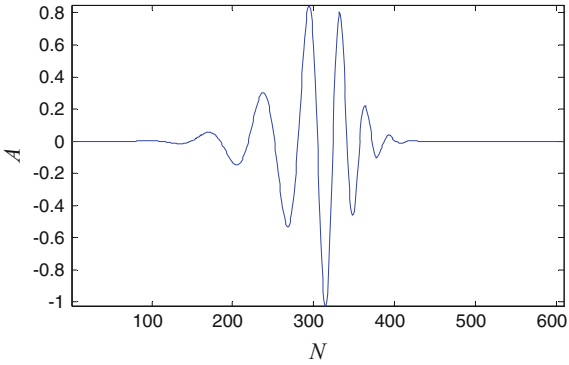


Fig. 3 The Db10 wavelet basis

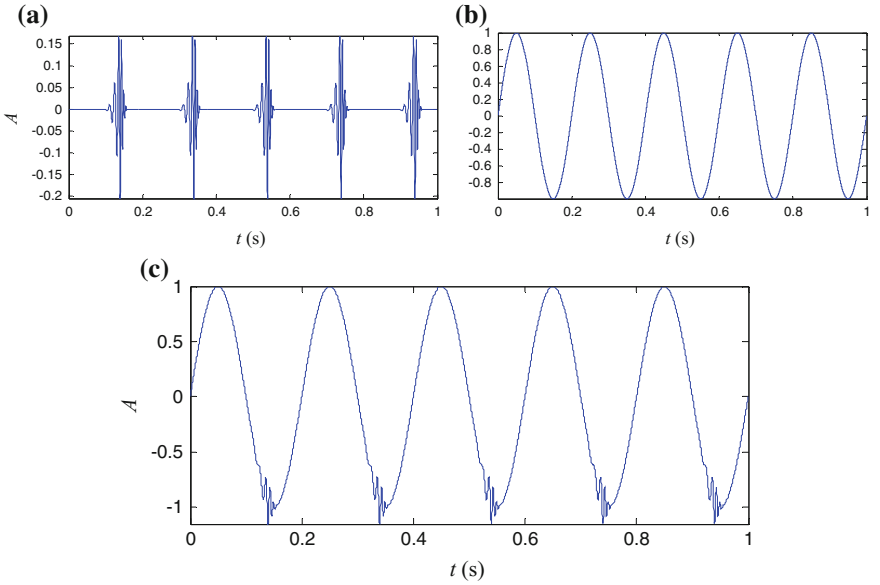


Fig. 4 **a** The periodic impulse component, **b** sinusoidal component and **c** the mixed simulated signal

four hundredth point to the one thousand and eighth point, which means that the signal begins from the 0.0781 s and lasts for $\tau = 0.2$ s.

First of all, He et al. [24] use the Db10 wavelet as the fundamental function to analyze the analog signal and the result showed in Fig. 5. As shown in Fig. 5, the signals from d1 to d5 are detail signals but the signals a5 is a approximation signal. During the experiment we often ignore the boundary effect of wavelet transform and it is clear that the periodic impulse component exists in the detail signals but the approximation signal a5 is exactly the sinusoidal component. It is natural that we use the detail signals d1 to d5 to construct the periodic impulse component and the

result shown in the Fig. 6a one more time. Afterwards we intercept the first factor of the reconstructed periodic impulse component and show it in Fig. 6b. Compared with the Db10 wavelet function in Fig. 3, we find it obvious that the first impulse element analyzed by the Db10 wavelet is absolutely similar to the real waveform of the impulse component. The correlation analysis shows the same result simultaneously, because it is 0.9969 that we finally got, and the correlation coefficient

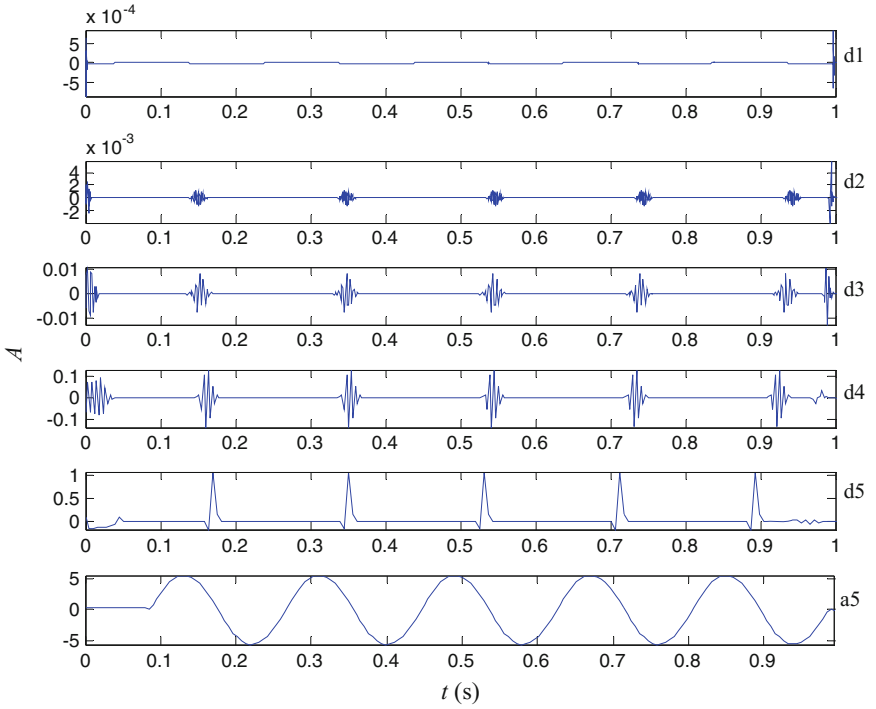


Fig. 5 The analyzed result by Db10 wavelet

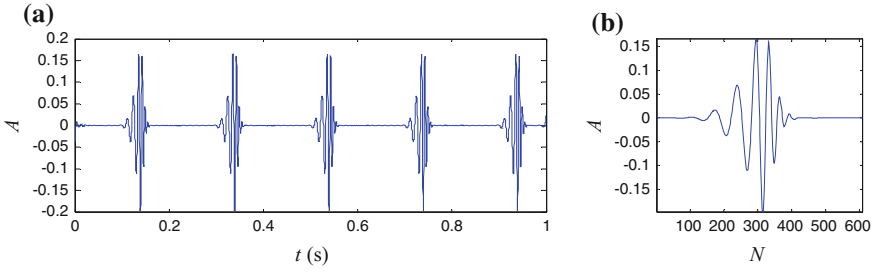


Fig. 6 **a** The reconstructed signal by Db10 wavelet and **b** the extracted impulse component

between the impulse element in the Fig. 6b and the impulse element in Fig. 3 is 0.9969, they are very close to 1.

$$\begin{cases} x(t) = 0.2x_1(t) + x_2(t) \\ x_1(t) = \sum_{i=1}^{5} \psi(t - 0.0781 - i\tau) \\ x_2(t) = \sin(10\pi t) \end{cases} \tag{7}$$

He et al. [24] also use other wavelet functions to justify the inner product theory, for example, the Db4 wavelet function, the Db32 wavelet function, the Sym10 wavelet function and the Bior3.7 wavelet function. The correlation coefficient between the extracted impulse component by Db4 in Fig. 7b and Db10 wavelet basis in Fig. 3 is calculated at 0.9319, which shows a worse result than Db10 wavelet. The symbol of A in Figs. 7 and 8 means the rising trend and B downtrend. The correlation coefficient between the extracted impulse component by Db32 in Fig. 8b and Db10 wavelet basis in Fig. 3 is calculated at 0.9570, showing that the

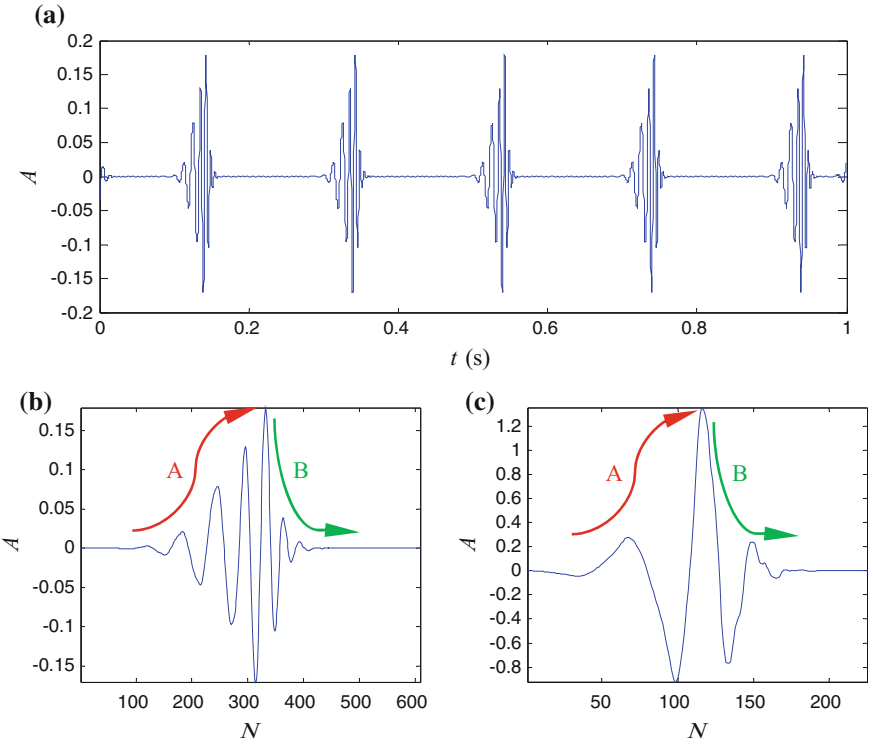


Fig. 7 **a** The reconstructed signal by Db4 wavelet, **b** the extracted impulse component and **c** the Db4 wavelet basis

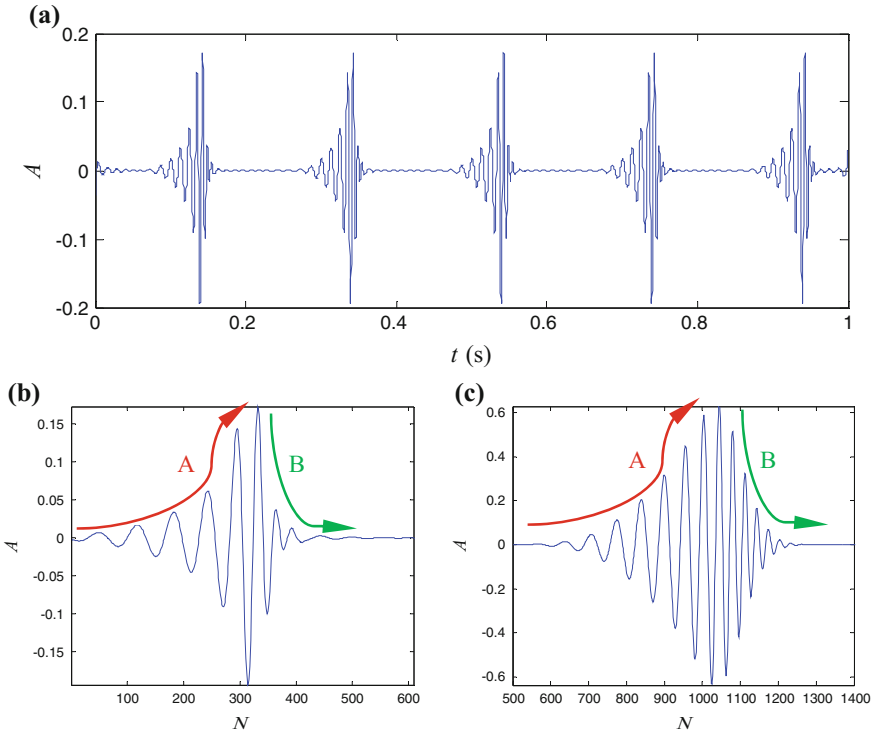


Fig. 8 **a** The reconstructed signal by Db32 wavelet, **b** the extracted impulse component and **c** the Db32 wavelet basis

Db32 wavelet is a better basis function than the Db4 wavelet. The Sym10 wavelet function is also used to analyze the simulated signals, and the correlation coefficient of the two impulse components in Figs. 9b and 3 is 0.9794. In the end, the Bior3.7 wavelet function is applied to the experiment, and the result of the coherence analysis between the impulse components in Figs. 10b and 3 is 0.8829. The mentioned result based on the above wavelet basis function is shown in Table 1.

According to simulation experiments and results, He et al. [24] find the following facts as: the correlation coefficient between Db10 wavelet transform coefficients with the impact element is the highest value (0.9969), close to 1. This value indicates Db10 wavelet has the highest similarity with emulational impulse component. And the extracted impulse component in Fig. 6b is almost identical with emulational impulse component. That is to say, choosing an appropriate wavelet is a very key step for the transform result. Moreover, an inappropriate wavelet employed in application will directly influence the accuracy of the signal feature extraction. These simulation validation results directly reveal the essence on inner product operation of WT.

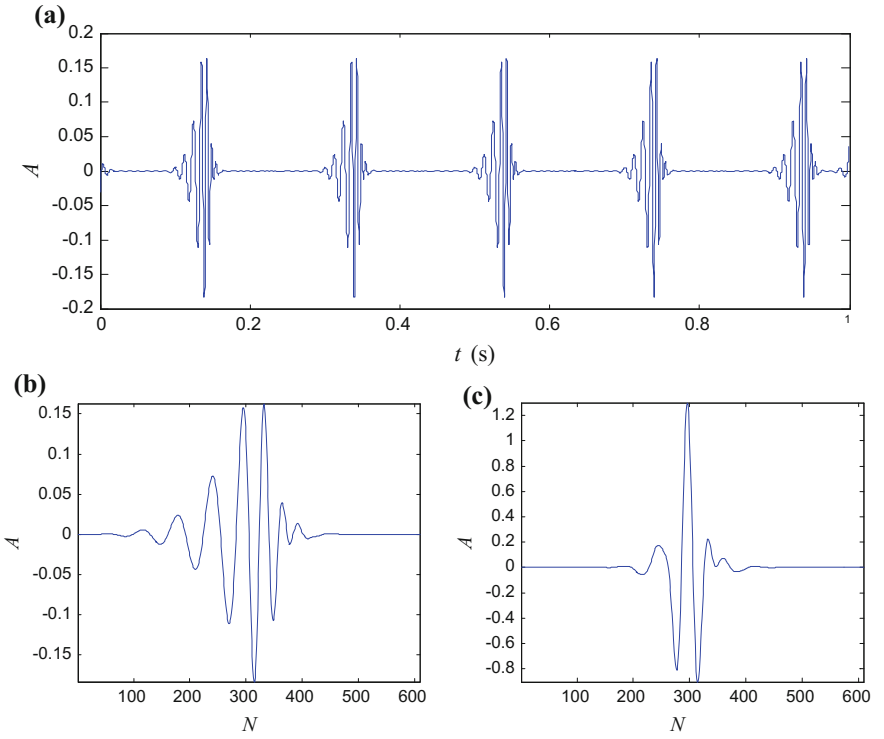


Fig. 9 **a** The reconstructed signal by Sym10 wavelet, **b** the extracted impulse component and **c** the Sym10 wavelet basis

What's more, He et al. [24] used the family of the Daubechies wavelets (DbN, N = 1–40 represents the different wavelet order, and the relevant wavelet function's vanishing moments is N) to analyze the analog signals synthetically. To assess the contribution of different wavelet functions, we used the correlation coefficient between the impulse component extracted by the diverse wavelet function and the emulational impulse component. The result of the Daubechies wavelet function is shown in the Table 2.

On the basis of simulation experiments and the results of them, He et al. [24] find some facts as follows: the correlation coefficient between the impact element with Db10 wavelet transform coefficients is the highest value (0.9969), mostly close to 1. The value indicates that emulational impulse component has the highest similarity with Db10 wavelet. And the impulse component that is extracted in Fig. 6b is nearly same as emulational impulse component. In other words, it is a very key step to choose an appropriate wavelet for the transform result. What's more, the accuracy of the extraction of signal feature will been influenced directly by an inappropriate wavelet that employed in application. These results of simulation validation directly reveal the nature of inner product operation of WT.

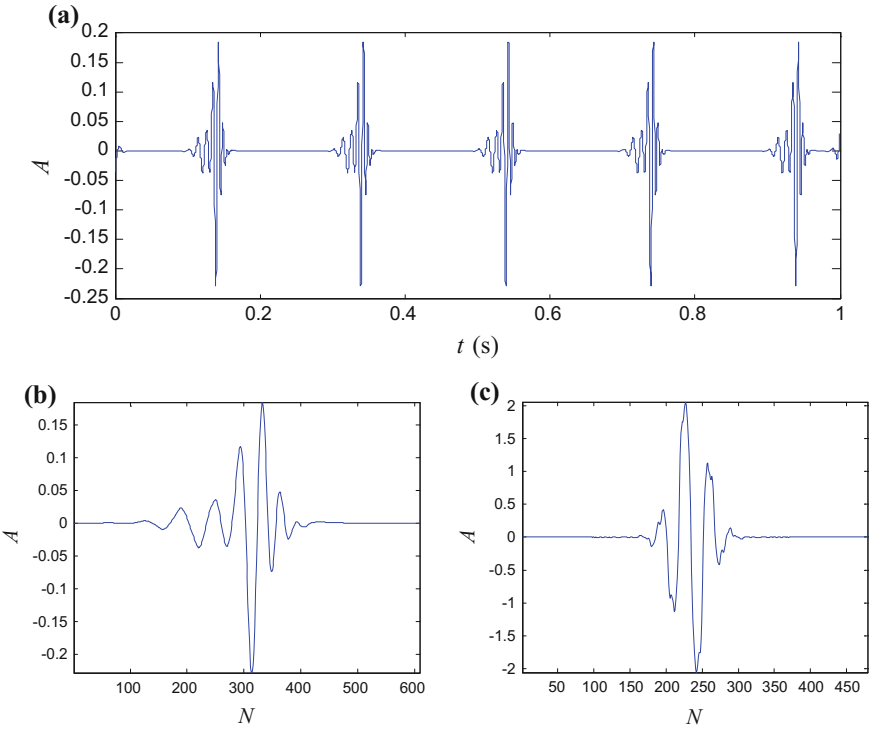


Fig. 10 **a** The reconstructed signal by Bior3.7 wavelet, **b** the extracted impulse component and **c** the Bior3.7 wavelet basis

Table 1 The coherence analysis results based on several typical wavelet basis

Wavelet	Db10	Db4	Db32	Sym10	Bior3.7
Correlation coefficient	0.9969	0.9319	0.9570	0.9794	0.8829

3 Adaptive Multiwavelet for RMFD

3.1 Summary of Multiwavelet Theory

Multiwavelet is the new advance of wavelet theory, and the wavelet basis of multiwavelet are generated by more than one mother wavelets [37]. The rising multiwavelet is extended from vector scaling and wavelet functions [37]. There are three more crucial properties of multiwavelet comparing with scalar wavelets. The first is that more synthetic information can be get from signal with multiwavelet than scalar wavelets owing to the varieties of shapes for mixed features, and thus the MT is fit for the fault diagnosis of rotating machinery [38]. The second is that multiwavelet can possess many properties at the same time including orthogonality,

Table 2 The coherence analysis results based on Db*N* wavelet basis

Db*N*	Correlation coefficient	Db*N*	Correlation coefficient
1	0.6058	21	0.9732
2	0.9126	22	0.9327
3	0.9910	23	0.9701
4	0.9319	24	0.9939
5	0.9099	25	0.9496
6	0.9886	26	0.9444
7	0.9719	27	0.9897
8	0.9118	28	0.9776
9	0.9638	29	0.9386
10	0.9969	30	0.9663
11	0.9353	31	0.9932
12	0.9351	32	0.9570
13	0.9958	33	0.9446
14	0.9703	34	0.9849
15	0.9248	35	0.9820
16	0.9713	36	0.9442
17	0.9948	37	0.9617
18	0.9421	38	0.9918
19	0.9425	39	0.9642
20	0.9937	40	0.9450

symmetry, compact support and higher order of vanishing moments, which cannot be satisfied simultaneously for scalar wavelet except Haar wavelet [38]. The last one is that there is higher freedom degree during the processing of constructing the multiwavelet basis [38].

The expression of multiwavelet could be $\Psi = (\psi_1, \ldots \psi_r)^T$, where T is the transpose. And the subspaces generating from a scaling function $\Phi = (\varphi_1 \ldots \varphi_r)^T$ can be expresses as

$$V_j = \overline{span(2^{j/2}\varphi_i(2^j t - k) : 1 \le i \le r, k \in Z)}$$

by its translation and dilation. Its complementary subspaces W_j

$$W_j = \overline{span(2^{j/2}\psi_i(2^j t - k) : 1 \le i \le r, k \in Z)},$$

is generated through the translation and dilation of multiwavelet. Similarly, the two-scale refinement Equations in the scalar case are still satisfied,

$$\Phi(t) = \sum_k H_k \Phi(2t - k) \tag{8}$$

$$\Psi(t) = \sum_k G_k \Phi(2t - k) \tag{9}$$

The recursive relationship of the coefficients $(c_{1,j,k}, c_{2,j,k})^T$ and $(d_{1,j,k}, d_{2,j,k})^T$ can be got through the dilation of Eqs. (8) and (9),

$$\begin{pmatrix} c_{1,j-1,k} \\ c_{2,j-1,k} \end{pmatrix} = \sqrt{2} \sum_{n=0}^{k} H_n \begin{pmatrix} c_{1,j-1,2k+n} \\ c_{2,j-1,2k+n} \end{pmatrix}, \quad j,k \in Z \tag{10}$$

$$\begin{pmatrix} d_{1,j-1,k} \\ d_{2,j-1,k} \end{pmatrix} = \sqrt{2} \sum_{n=0}^{k} G_n \begin{pmatrix} c_{1,j-1,2k+n} \\ c_{2,j-1,2k+n} \end{pmatrix}, \quad j,k \in Z \tag{11}$$

Furthermore,

$$\begin{pmatrix} c_{1,j,n} \\ c_{2,j,n} \end{pmatrix} = \sqrt{2} \sum_{n=0}^{k} \left(H_k^+ \begin{pmatrix} c_{1,j-1,2k+n} \\ c_{2,j-1,2k+n} \end{pmatrix} + G_k^+ \begin{pmatrix} d_{1,j-1,2k+n} \\ d_{2,j-1,2k+n} \end{pmatrix} \right) \tag{12}$$

The input streams of multiwavelet differs from scalar wavelets because of matrix-valued filter-bank, and that means both the process of decomposition and reconstruction are multi-input and multi-output (MIMO). Therefore, the pre-process of the input signal avant decomposition and the post-process of the output signal after reconstruction is inevitable [39]. Post-filter and pre-filter are the inverse matrixes for each other. There exist two kinds of pre-processing method [39]. The one on the base of an oversampling scheme is repeated row pre-filter [39]. And the other is on the base of critical sampling scheme [39]. It has been proved that oversampling during feature extraction is effective [39]. For the reason that the first one leading to double oversampling of the data, usually in engineering applications, people adopt the repeated row pre-filter. So the decomposition and reconstruction above can be expressed in Fig. 11. Where $Q(\omega)$ represents the pre-filter, and $P(\omega)$ represents the post-filter.

In order to utilize the useful information existing in high frequency band, which may be left out because that multiwavelet transform merely concentrates on multi-resolution analysis in low frequency band, the multiwavelet packet is developed.

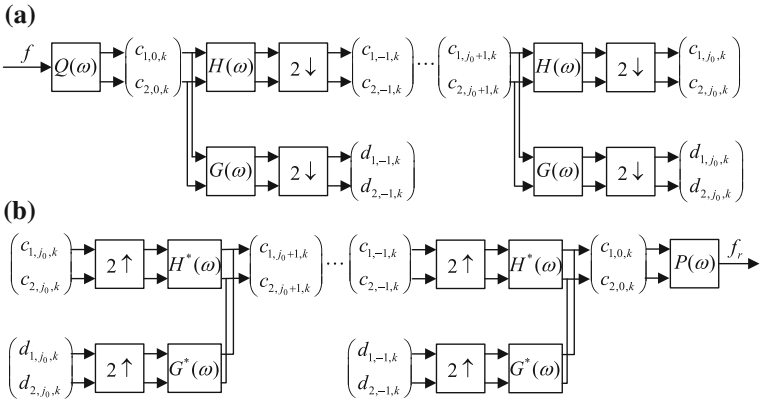


Fig. 11 The two-stage decomposition and reconstruction processes of MT

3.2 *Adaptive Multiwavelet Construction*

In order to construct new biorthogonal wavelets, the lifting scheme (LS) is proposed [16, 40]. Because that there is more designing freedom for multiwavelet than scalar ones, the symmetric lifting (SymLift) scheme is adopted to construct new multiwavelet.

If ω_0 is used as an origin to produce a new wavelet with the assigned numbers of vanishing moments, accordingly lifting scheme could be

$$\omega_0^{new} = \omega_0(x) + \sum_{i=1}^{k} \lambda_i \omega_i(x) \tag{13}$$

where λ_i is the lifting coefficients. For satisfying symmetry, the major factor is selecting the translation quantity τ of functions. Let's take the symmetric lifting of Ψ_1 for example, assuming that ω_i is symmetric/ antisymmetric functions about the points a_{ω_i} separately. The translation quantity τ must meet the Equation below,

$$a_{\Psi_1} - (a_{\omega_i} + \tau_{\omega_i,1}) = (a_{\omega_i} + \tau_{\omega_i,2}) - a_{\Psi_1} \text{ and } \left|\tau_{\omega_i,1}\right| = \left|\tau_{\omega_i,2}\right| \tag{14}$$

where $i = 1, 2$. Let $B_{\omega_i} = \pm 1$ (1 represents symmetry and -1 represents anti-symmetry), means the symmetry/antisymmetry properties of the original multiwavelet. Lifting vanishing moment p of Ψ to p' , a linear Equation system can be got through integrating both sides of Eq. (13).

$$\begin{bmatrix} \int \omega_1(x+\tau_{\omega_1,1})x^p dx & \int \omega_1(x+\tau_{\omega_1,2})x^p dx & \cdots \\ \int \omega_1(x+\tau_{\omega_1,1})x^{p+1} dx & \int \omega_1(x+\tau_{\omega_1,2})x^{p+1} dx & \cdots \\ \vdots & \vdots & \vdots \\ \int \omega_1(x+\tau_{\omega_1,1})x^{p'-1} dx & \int \omega_1(x+\tau_{\omega_1,2})x^{p'-1} dx & \cdots \end{bmatrix} \begin{bmatrix} 1 & & 0 \\ B_{\Psi_i} B_{\Phi_i} & & \\ & \ddots & \\ 0 & & \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} -\int \omega_0(x)x^p dx \\ -\int \omega_0(x)x^{p+1} dx \\ \vdots \\ -\int \omega_0(x)x^{p'-1} dx \end{bmatrix} \tag{15}$$

The lifting coefficients of Ψ_1 are the results of Eq. (15). And the lifting of Ψ_2 is similar. Then, substitute the lifting coefficients into Eq. (13). After that, acquire the lifting matrices T and S using Z transform. The detail could be expressed as below,

$$G_{new}(z) = T(z^2)[G(z) + S(z^2)H(z)] \tag{16}$$

With $T(z)$ and $S(z)$, the construction of a new symmetric biorthogonal multiwavelet can be done, where $T(z)$ and $S(z)$ are finite order and determinant of $T(z)$ is monomial.

The only case that there are many solutions for c_i is an underdetermined linear Equation system thinking back to the SymLift scheme for multiwavelet in Eq. (15). That means there are many alternatives of multiwavelet with a assigned increase of vanishing moment. Thus, there exist a lot of additional designing freedoms when constructing multiwavelet. Equation (15) could be simplified as $MB\lambda = N$. Assuming the vanishing moments of the multiwavelet after SymLift is p' and Rank$(\cdot)$ means the rank of a matrix.

$$N_f = (p' - 1) - \text{Rank(MB)} \tag{17}$$

And, free parameters for the design of the new biorthogonal multiwavelet exist.

In order to match and model the fault features, the best parameter relied on the input dynamic response signals need to be selected carefully according to an evaluation rule. It is famous that Shannon entropy is an effective rule to represent the diversity of possibility distribution. Therefore, the minimum entropy principle is promoted to search for the best parameters through measuring the sparsity. The recommended tool to optimize multiwavelet having free parameters is Genetic algorithm (GA). In this chapter, the searching ranges of these parameters are consistently within the interval $[-50, 0) \cup (0, 50]$. In order to accelerate the process of optimizing GA, the population scale is assigned 100 and the number of iterations is assigned 50, the possibility of crossover is assigned 0.6 and the possibility of mutation is assigned 0.05.

To display the synthetical information about the consequences of redundant multiwavelet packet decomposition, the ratio r representing relative energy of the interested characteristic frequency is calculated to choose the sensitive frequency band.

$$r = \frac{\{\max[A(f_c)]\}^2}{\sum_{f=0}^{f'} A(f)^2}, \quad f_c \in (f_c - \Delta, f_c + \Delta) \tag{18}$$

where A means the amplitude of envelope spectrum of redundant multiwavelet packet coefficients, f represents the characteristic frequency and Δ means the frequency interval. $f = 0 \sim f'$ represents the range of the frequency band. The sensitive range could be chosen by the comparatively larger frequency band, which can judge whether the machinery fault exists. The step of the adaptive multiwavelet for RMFD is expressed as Fig. 12.

3.3 Experimental Study

In order to simulate the many kinds of abnormal mechanical phenomena occurring on the components, a multifunctional double-motor transmission train rig without clearance was designed, as Fig. 13 shows. The Fig. 13a describes the schematic diagram of the test rig, and the Fig. 13b display the rig.

The experiment simulates mixed-fault of a bevel gearbox including a rubbing fault on the spiral bevel gear and a bearing fault on the outer-race. The installation of fault pieces is shown in Fig. 13a red section. The fault test specimens are displayed in Fig. 14.

The number of gearwheel teeth in test gearbox is shown in Table 3. The fault gear locates in the input terminal of the gearbox is a spiral bevel gear. Experiment bearing is 32908, whose structure parameters are displayed in Table 4. The reduction ratio of planetary gearbox is 32 and the motor's rotate speed is 1500 r/min, the characteristic frequencies are shown in Table 5. With Eqs. (19–21) and the parameters of the bearing in Table 2, the character frequency of bearing can be calculated, as displayed in Table 3.

$$f_e = \frac{D}{2d} \cdot f_r \cdot \left(1 - \left(\frac{d}{D}\right)^2 \cos^2\theta\right) \tag{19}$$

$$f_o = \frac{1}{2} \cdot N \cdot f_r \cdot \left(1 - \frac{d}{D}\cos\theta\right) \tag{20}$$

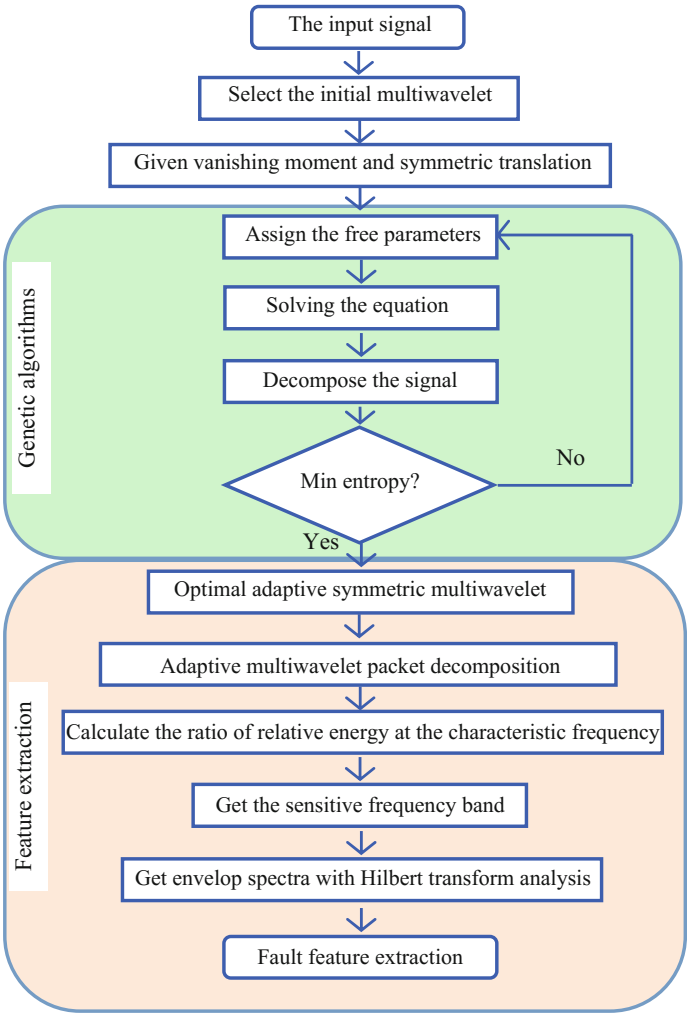


Fig. 12 The flow chart of the adaptive multiwavelet method for RMFD

$$f_i = \frac{1}{2} \cdot N \cdot f_r \cdot \left(1 + \frac{d}{D} \cos\theta\right) \tag{21}$$

where f_e, f_o and f_i represent the characteristic frequencies of roller element, outer-race and inner-race of bearing, respectively. The parameter f_r means the rotation frequency, D means the pitch diameter, d means the roller diameter, N means the number of roller, and θ means the contact angle.

There are vibration acceleration sensors installed in the output terminal of planetary gearbox as shown in Fig. 20b, and the vibration signal of both horizontal

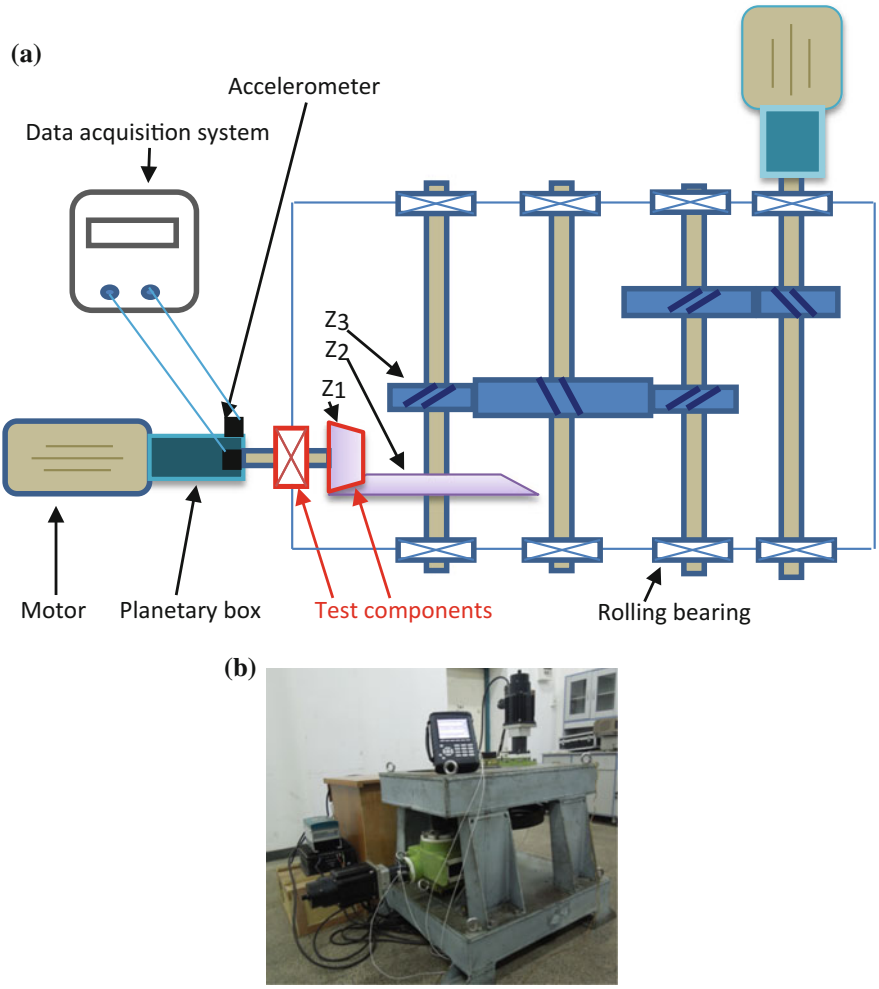


Fig. 13 The experimental setup. **a** The schematic diagram and **b** the test rig

and vertical can be obtained. The sample frequency is 5.12 kHz, motor speed is 1414 rpm and the sample time is 30 s. Figure 15a shows the original signal of the vertical sensor, showing severe background noise with the impulse information being concealed. Figure 15b shows the FFT spectrum of the vibration and Fig. 15c displays the original Hilbert envelope spectrum. Figure 15d shows the amplified spectrum of the original signal. There are only the gear meshing frequency (7 Hz) and its doubling frequency in the amplification of the low frequency band. It is challenging to extract spiral bevel gear's fault features for that the contemporary contact of a few teeth and stable transmission lead to large overlap ratio. Thus, it has been one of the most challenging tasks to do the fault diagnosis of spiral bevel gear.

Fig. 14 The defects of the testing components

Table 3 Parameters of the gears in the test gearbox

Pinion wheel teeth number (Z_1)	Gearwheel teeth number (Z_2)	Output stage wheel teeth number (Z_3)
19	38	19

Table 4 Parameters of the test rolling bearing

Outer race diameter (mm)	Inner race diameter (mm)	Roller diameter mean value (mm)	Roller number	Contact angle (°)
62	40	4.635	26	13° 50′

Table 5 The feature frequencies of the gearbox in case 1

Feature frequency	Value (Hz)
Gear meshing frequency of the fist stage planet gearbox	247.5
Gear meshing frequency of the second stage planet gearbox	61.875
Gear meshing frequency of the spiral bevel gear	13.995
Gear meshing frequency of the output stage gear	6.9978
Feature frequency of pinion wheel	0.7366
Characteristic frequency of inner-race fault of rolling bearing (32908)	10.42
Characteristic frequency of outer-race fault of rolling bearing (32908)	8.73
Characteristic frequency of roller fault of rolling bearing (32908)	4.02

In order to acquire the whole fault information, the original signal will be analysed using this method. The original multiwavelet chosen is cubic Hermite and the vanishing moment is lifted to 5, moreover, the multiwavelet lifted is displayed in Fig. 16. Figure 17 shows the histogram of relative energy ratio when three layers redundant multiwavelet packet decomposition has been done.

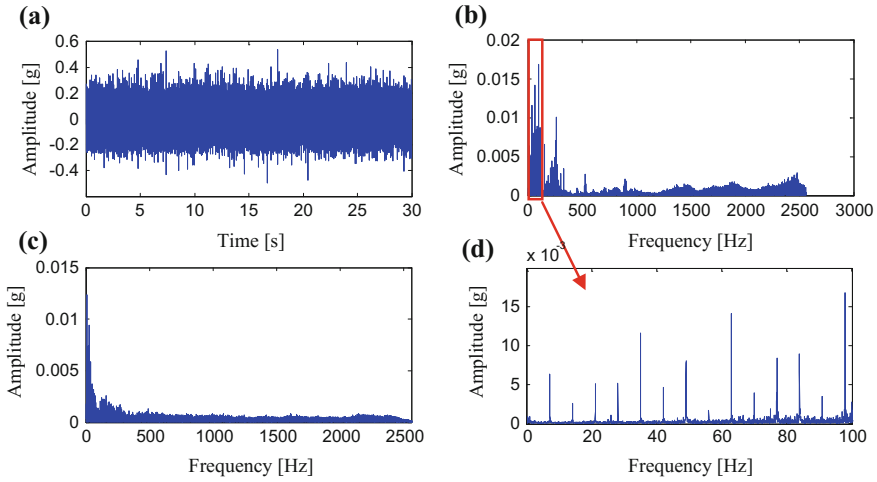


Fig. 15 **a** The vibration signal of the test in the case 1, **b** the FFT spectrum, **c** the envelop spectrum, **d** zoomed-in (**b**)

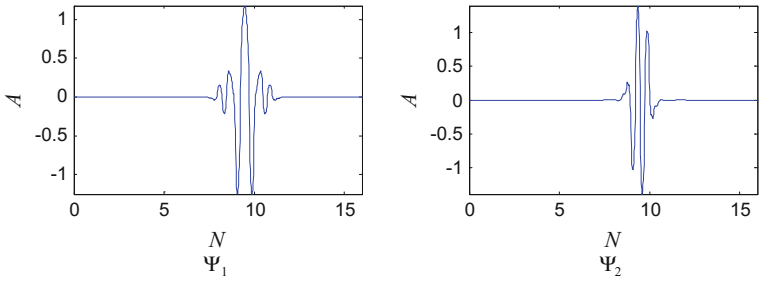


Fig. 16 Multiwavelet functions of the adaptive multiwavelet

It can be seen that the faults of spiral bevel gear and outer-race bearing are more distinct and the faults of inner-race bearing and roller element may be weak from Fig. 17. Meanwhile, the highest relative energy ratio of the faults of spiral bevel gear, roller element and inner-race bearing are in the 6th branch and the highest relative energy ratio of the faults of outer-race bearing is in the 9th branch. When the 6th branch and 9th branch are selected as sensitive frequency bands, Fig. 18 shows the result of envelope spectra. Here, the spectrum of 0.733 Hz same as the spiral bevel gear's characteristic frequency is clear. Base on that, the early fault of the spiral bevel gear can be diagnosed and the characteristic frequency of the roller element and the inner-race bearing are submerged in background noise. Though it's not the dominant in this branch, outer-race bearing fault is also here. The obvious 8.73 Hz in Fig. 18b indicates the bearing's outer-race fault. The spiral bevel gear's characteristic frequency exists in this branch. Though there is strong background

noise in this experiment leading to extreme difficulties, the spiral bevel gear and outer-race bearing's unobvious mixed-fault features could be extracted with the construction and analysis of the proposed method. In order to verify the effectiveness of the proposed method, Db8 has been used to analyse the signal, which is shown in Fig. 19.

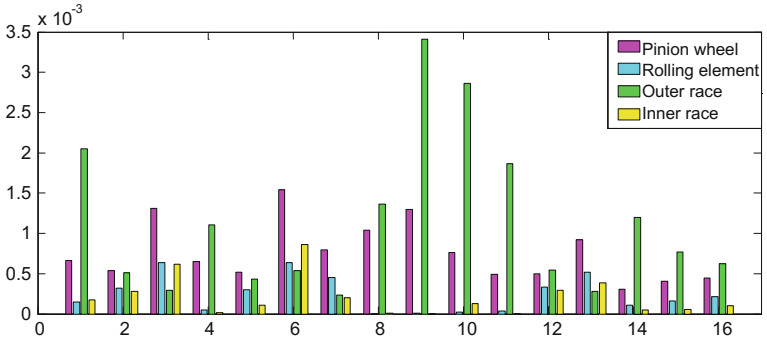


Fig. 17 The ratios on the sixteen multiwavelet packet coefficients in the case 1

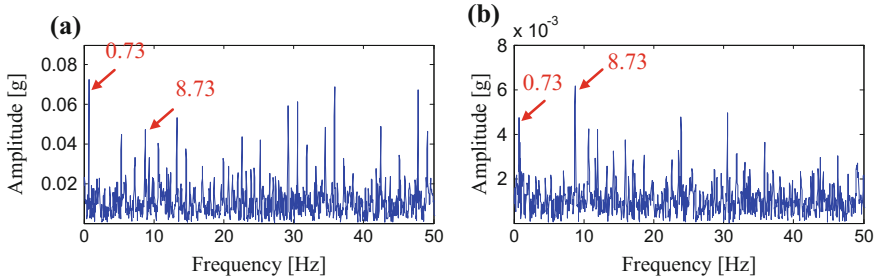


Fig. 18 The envelope spectrums of two sensitive frequency bands with the maximum ratio of the different components in case 1. **a** The envelope spectrum of the 6th branch, **b** the envelope spectrum of the 9th branch

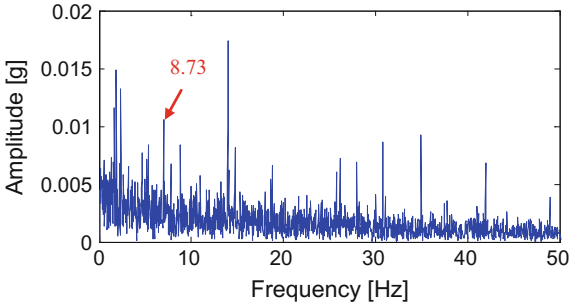


Fig. 19 The analyzed result of the vibration signal in case 1 using redundant wavelet packet transform by Db8 with envelop spectrum

It can be seen that all these methods can diagnose the outer-race bearing fault, but the spiral bevel gear's fault cannot be successfully extracted. The major reason is spiral bevel gear's overlap ratio. Thus, the proposed method in this chapter can be effective in diagnosing the bearing and gear mixed-fault.

4 Discussion

According to the descriptions above in this chapter, it can be seen that there are many successful applications in RMFD with wavelet transform. But it's more difficult to realize a standard status in engineering applications than FT due to the abundance of basis functions selection as well as varieties of transform schemes. These crucial factors lead to the current status of wavelet in fault diagnostics as well as doubt on wavelet's engineering applications from field staff. The fact is that while researchers appreciate wavelet transform for its capability, field staff dislike it due to its comparatively complicated steps. Therefore, basing on the inner product for RMFD, super wavelet transform (SWT) and more rational threshold shrinkage strategy need to be studied in the future. It can be seen from Fig. 20 that there are

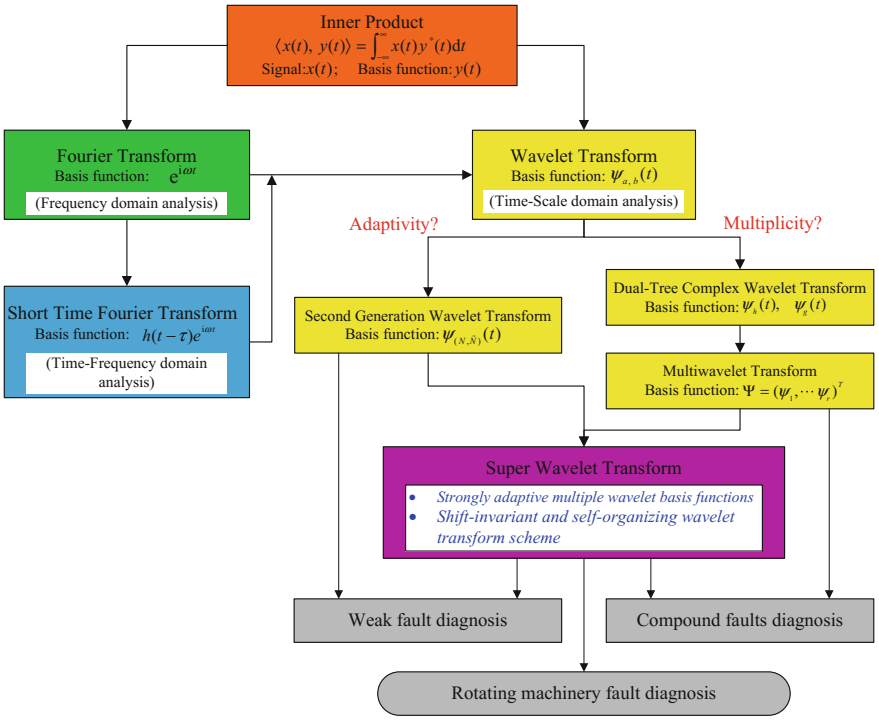


Fig. 20 The developments process of WT based on inner product for RMFD

two major factors restrict the capability of super wavelet, namely strongly adaptive multiple wavelet basis functions and shift-invariant self-organizing wavelet transform scheme, and both the two factors can help to improve the current status in engineering applications. What's more, research on more rational data-driven threshold shrinkage strategy used in wavelet threshold denoising (WThD) [41, 42] is helpful to improve the capability of WT in RMFD with heavy background noise.

5 Conclusion

In this chapter, simulation experiments are introduced first to validate the inner product essence of WT in RMFD. Then the development of WT on the base of inner product is concluded and the relevant applications in RMFD are summed up. What's more, the construction and applications of key developments on adaptive multiwavelet in RMFD are also recommended. Eventually, super wavelet transform is presented and discussed as a crucial prospect of WT based on inner product. Looking forward to that this chapter can synthesize individual pieces of information about WT-based fault diagnosis and provide an in-depth and synthetical references for relative researchers to help them to find out further research topics in this field.

References

1. Feng Z.P., Zuo M.J., "Fault diagnosis of planetary gearboxes via torsional vibration signal analysis," Mechanical Systems and Signal Processing, 2013, 36:401–421.
2. Li J.M., Chen X.F., He Z.J., "Multi-stable stochastic resonance and its application research on mechanical fault diagnosis," Journal of Sound and Vibration, 2013, 332(22):5999–6015.
3. Marquez F., Tobias A., Perez J., et al., "Condition monitoring of wind turbines: techniques and methods," Renewable Energy, 2012, 46:169–178.
4. Chen B.Q., Zhang Z.S., Zi Y.Y., et al., "Detecting of transient vibration signatures using an improved fast spatial–spectral ensemble kurtosis kurtogram and its applications to mechanical signature analysis of short duration data from rotating machinery," Mechanical Systems and Signal Processing, 2013, 40:1–37.
5. Sun H.L., He Z.J., Zi Y.Y., et al., "Multiwavelet transform and its applications in mechanical fault diagnosis – A review, Mechanical Systems Signal Processing," 2014, 43(1–2):1–24.
6. Qin, S.R., Zhong Y.M., 2004 "Research on the unified mathematical model for FT, STFT and WT and its applications," Mechanical Systems and Signal Processing, 18 (6):1335–1347.
7. Nilsen G.K., "Recursive Time-Frequency Reassignment, IEEE Transactions on Signal Processing," 2009, 57(8):3283–3287.
8. Liu X.P., Shi J., Sha X.J., et al., "A general framework for sampling and reconstruction in function spaces associated with fractional Fourier transform," Signal Processing, 2015, 107:319–326.
9. Giv H.H., "Directional short-time Fourier transform," Journal of Mathematical Analysis and Applications, 2013, 399(1):100–107.

10. Baccar D., Söffker D., "Wear detection by means of wavelet-based acoustic emission analysis," Mechanical Systems Signal Processing, 2015, 60–61:198–207.
11. Ł. Jedliński, J. Jonak, "Early fault detection in gearboxes based on support vector machines and multilayer perceptron with a continuous wavelet transform," Applied Soft Computing, 2015, 30: 636–641.
12. Gao R.X., Yan R.Q., Wavelets: Theory and Applications for Manufacturing, Springer-Verlag, 2011.
13. Randall R.B., Vibration-based Condition Monitoring: Industrial, Automotive and Aerospace Applications, Wiley, 2011.
14. Lei Y.G., Lin J., Zuo M. J., et al., "Condition monitoring and fault diagnosis of planetary gearboxes: A review," Measurement, 2014, 48:292–305.
15. Li B., Chen X.F., "Wavelet-based numerical analysis: A review and classification." Finite Elements in Analysis and Design, 2014, 81:14–31.
16. Sweldens W., "The lifting scheme: A custom design construction of biorthogonal wavelets," Applied Computational Harmonic Analysis, 1996, 2:186–200.
17. Li Z., He Z.J., Zi Y.Y., et al., "Rotating machinery fault diagnosis using signal-adapted lifting scheme," Mechanical Systems Signal Processing, 2008, 22(3): 542–556.
18. Xiao W.R., Zi Y.Y., Chen B.Q., et al., "A novel approach to machining condition monitoring of deep hole boring," International Journal of Machine Tools and Manufacture, 2014, 77: 27–33.
19. Kingsbury N.G., "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters," IEEE Digital Signal Processing Workshop, 1998.
20. Donovan G.C., Geronimo J.S., Hardin D.P., et al., "Construction of orthogonal wavelets using fractal interpolation functions," SIAM J. on Mathematical Analysis, 1996, 27(3): 1158–1192.
21. Zfian, M.H. Moradi, S. Gharibzadeh, "Microarray image enhancement by denoising using decimated and undecimated multiwavelet transforms," Signal Image and Video Processing, 2009, 4:177–185.
22. Downie T.R., Silverman B.W., "The discrete multiple wavelet transform and thresholding methods," IEEE Trans. on Signal Processing, 1998, 46:2558–2561.
23. Peng Z.K., Chu F.L., "Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography," Mechanical Systems Signal Processing, 2004, 18:199–221.
24. He Z.J., Zi Y.Y., Chen X.F., "Transform principle of inner product for fault diagnosis". Journal of Vibration Engineering, 2007, 20(5):528–533.
25. B.Q. Chen, Z.S. Zhang, C. Sun, et al., "Fault feature extraction of gearbox by using overcomplete rational dilation discrete wavelet transform on signals measured from vibration sensors," Mechanical Systems Signal Processing, 2012, 33:275–298.
26. Mallat S., A Wavelet Tour of Signal Processing, Second Edition, Academic Press, 2003.
27. Lin J., Qu L.S., "Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis," Journal of Sound and Vibration, 2000, 234(1):135–148.
28. Lin J., Zuo M.J., "Gearbox fault diagnosis using adaptive wavelet filter," Mechanical Systems and Signal Processing, 2003, 17(6):1259–1269.
29. Grossmann A., Morlet, J., "Decomposition of hardy functions into square integrable wavelets of constant shape," SIAM Journal on Mathematical Analysis, 1984, 15(4):723–746.
30. Meyer Y. "Principe d'incertitude, bases hilbertiennes et algebres d'operateurs," Séminaire Bourbaki, 1986, 662:209–223.
31. Mallat S., "Multiresolution approximations and wavelet orthonormal bases of L2(R)," Transactions of the American Mathematical Society, 1989, 315(1):69–87.
32. Mallat S., "A theory for multiresolution signal decomposition - The wavelet representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989 11(7):674–693.
33. Daubechies I., "Orthonormal bases of compactly supported wavelets," Communications on Pure and Applied Mathematics, 1988, 41(7):909–996.

34. Daubechies I., “Ten lectures of wavelets,” CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics: Philadelphia, PA, 1992.
35. Coifman R.R., Meyer Y., Quake S., et al., “Signal processing and compression with wavelet packets,” Wavelets and Their Applications, 1994, 442:363–379.
36. Li H., Zhang Y., Zheng H., “Application of Hermitian wavelet to crack fault detection in gearbox,” Mechanical Systems and Signal Processing, 2011, 25(4):1353–1363.
37. Keinert F., Wavelets and Multiwavelets, Studies in Advanced Mathematics, Chapman & Hall/CRC Press, 2004.
38. Alpert B., “A class of basis in L2 for the sparer representation of integral operators,” SIAM Journal on Mathematical Analysis, 1993, 24 (1):246–262.
39. Jiang Q.T., “Orthogonal multiwavelet with optimum time-frequency resolution,” IEEE Transactions on Signal Processing, 1998, 46(4):830–844.
40. Wang X.D., Zi Y.Y., and He Z.J., “Multiwavelet denoising with improved neighboring coefficients for application on rolling bearing fault diagnosis,” Mechanical Systems and Signal Processing, 2011, 25:285–304.
41. Cai T.T., Silverman B.W., “Incorporating information on neighboring coefficients in wavelet estimation,” Sankhya Series B, 2011, 63:127–148.
42. Chen J.L., Li Z.P., Pan J., Chen G.G., Zi Y.Y., and Z.J. He, “Wavelet transform based on inner product in fault diagnosis of rotating machinery: A review,” Mechanical Systems and Signal Processing, 2016, 70–71:1-35.

Wavelet Based Spectral Kurtosis and Kurtogram: A Smart and Sparse Characterization of Impulsive Transient Vibration

Binqiang Chen, Wangpeng He and Nianyin Zeng

Abstract Mechanical signature analysis is of vital importance to the structural health monitoring of mechanical equipment. However, the fast development of mechanical signature analysis tool always requires a rich and deep understanding of state-of-the-art technologies, which is often lacked by the on-site staff. In this chapter, we introduce an effective methodology that ensure automatic detection of impulsive transient vibrations occurring during machinery fault events. This methodology is originally derived from the concept of spectral kurtosis, whose advent has a close relation with the early development of wavelet theory, and acquired a fast computation implementation named fast kurtogram. The essential originality of this methodology lies in its unique way of combining multi-scale analysis and scalar indicator based characterization of impulsive transient components. As a result, this methodology emerges as a single-input-single-output system for both theoretical researchers and on-site engineers. In the presented materials, basics and fundamentals of this fast developing methodology are introduced. The recent improvements mainly focus on the construction of new multi-scale signal decomposition frames and the invention of new scalar-valued indicators. All the efforts are motivated to obtain a satisfactory sparse characterization of impulsive transient components induced by machinery faults. A range of construction examples of wavelet-based spectral kurtosis with their engineering applications are presented to demonstrate the developments.

B. Chen (✉) · N. Zeng
School of Aerospace Engineering, Xiamen University,
361005 Xiamen, Fujian, People's Republic of China
e-mail: cbq@xmu.edu.cn

W. He
School of Aerospace Science and Technology, Xidian University,
710071 Xi'an, Shaanxi, People's Republic of China

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors,
Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_5

1 A Brief Introduction

Rotating machinery (RM) covers a wide range of applications in the industrial manufacturing field. Rotating machinery is composed of various types of rotating mechanical components such as shafts, rotors, roller element bearings and gear transmission. Structural health monitoring plays an important role to prevent machinery downtime and catastrophic events that result in loss of economic benefits as well as human lives. A variety methods, including vibration based condition monitoring, acoustic emission analysis, temperature trend analysis and wear debris analysis, have been developed for implementation of both diagnostic aspect and prognostic aspect during rotating machinery maintenance [1, 2], vibration based analysis has become the preferred mean to understand the operation state of rotating machinery. One of the fundamental premises of vibration-based condition monitoring is that the abnormal states of a machinery that lead to symptomatic vibrations that can be easily distinguished from a healthy reference [2, 3].

Owing to the existence of multiple types of mechanical components and their coupled rotating transmissions, the occurrence of mechanical faults will produce repetitive impulsive transient vibrations masked by other non-stationary vibrations and interfering noises. Therefore, a most important problem in diagnosis and prognosis of RM is concluded as the recovery of the incipient and critical vibration transients [4–7]. Considerable efforts have been paid to address this issue and fruitful achievements are continuously being acquired. Among the massive materials, a few famous examples are mentioned as STFT, continuous wavelet transform, discrete wavelet transform, empirical mode decomposition and other adaptive signal analysis tools. However, each of them needs to be combined with some specific pre-processing procedures and post-processing procedures to attain a concrete diagnostic result, which often requires sophisticated and professional human interference that are usually lacked by on-site staffs and industrial engineers. Owing to the present situation, the state-of-the-art techniques are not receiving good propagations.

Inspired by the difficult issue, spectral kurtosis (SK) and its fast computation implementation, Kurtogram, were investigated and have attracted considerable attentions. This methodology becomes very popular to the on-site vibration measurement due to its smart and powerful capability in detecting impulsive transients. Spectral kurtosis was proposed by Dwyer [6] as a spectral statistic which helpfully supplements the classical power spectral density. SK by Dwyer was initially formulated as a fourth-order moment of real part of short-time Fourier transform such that it can indicate non-Gaussian components in signal. However, this initial definition of SK was seldom utilized due to its complicated properties. In 1994, Otonnello and Pagnan proposed an improved SK based on normalized fourth-order moment of the magnitude of STFT [8, 9]. Their definition showed SK is capable of recovering randomly occurring signals severely corrupted by additive stationary noise. In 1996, with the mature establishment of higher-order statistics, more formal definitions of SK were given to enhance its filtering ability. Capdevielle considered

Table 1 Different detection filters used to estimate the SK of vibration signal

Detection filter	SK technique	References	Comments
CMWT	STFT-based SK	[15]	Uniform resolution on a logarithmic frequency scale
WPT	Kurtogram	[16]	More dedicated division of the time-frequency decomposition
Adaptive supper-position window in frequency domain		[17]	
TQWT		[18]	More flexible wavelet transform of tunable quality factor
QAWTF	Kurtogram	[4]	Quasi-analytic wavelet tight frames
AKBS	Adaptive spectral kurtosis	[19]	Remove sinusoidal interferences
Mulitwavelet transform	Kurtogram	[20]	Customized multiwavelet transform
Morlet wavelet	Adaptive spectral kurtosis	[21]	Morlet wavelet used as filterbank

SK as the normalized fourth-order cumulant of the Fourier transform, and used it as a measure of distance of a process form Gaussianity [10]. In 2006, Antoni provided a new SK definition in terms of Wold-Cramer decomposition and introduced many conditional non-stationary cases for which his definition can be applied [11].

Kurtogram [12] is usually referred to the fast computational implementation of spectral kurtosis [13]. The using of different kinds of detection filters will result in distinguished variations of kurtogram. The Short time Fourier transform (STFT) and multi-rate filterbank (MRFB) based quint wavelet (QW) were originally investigated and turn out to be efficient estimates of the impulsive transients. However, efforts were paid to enhance the effectiveness of kurtogram. These efforts can be categorized into two aspects. The first is the development of time-frequency frames, and the other is the developments of new statistical indicators with similar functions to SK. In this chapter we focus on the wavelet-based and enhanced spectral kurtosis techniques. Table 1 show the different detection filters used to replace the original ones of STFT and QW [14].

2 Spectral Kurtosis and Fast Kurtogram

2.1 Signal Modelling

An essential problem of spectral kurtosis is how to detect non-stationary fault signal $Y(t)$ in a vibration measurement $Z(t)$ in the presence of some strong additive noise

$N(t)$ as well as some other interfering components. This signal model is partly based on the work of Antoni and Randall [11–13].

$$Z(t) = Y(t) + N(t) \tag{1}$$

It is assumed that $Y(t)$ is of impulsive transient nature. In terms of Wold's decomposition, any stationary stochastic process can be treated as the output of a causal, linear, and time-invariant system $h(s)$ excited by strict white noise:

$$Y(t) = \int_{-\infty}^{t} h(t-\tau)X(\tau)d\tau \tag{2}$$

No restriction is imposed on $X(t)$ other than it has a flat spectrum almost everywhere and it has a symmetric probability density function. While its counterpart in the frequency domain is named as the Cramer's decomposition via

$$Y(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{j2\pi ft} H(f)d\widehat{X}(f) \tag{3}$$

where $H(f)$ is the Fourier transform of $h(s)$ and $d\widehat{X}(f)$ is the spectral counterpart associated with $X(t)$

$$X(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{j2\pi ft} d\widehat{X}(f) \tag{4}$$

Different interpretation may be applied to Eq. (6). A famous one is regarded as the integrant is the filtering of $Y(t)$ with an infinitely narrow-band filter cantered at the frequency of f. In more general cases, the transfer function $h(s)$ can be time-varying, such that an extended Wold-Cramer decomposition of non-stationary process can be represented as

$$Y(t) = \int_{-\infty}^{t} h(t, t-\tau)X(\tau)d\tau \tag{5}$$

while the spectral counterpart of Eq. (5) is

$$Y(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{j2\pi ft} H(t,f)d\widehat{X}(f) \tag{6}$$

The understanding of Eq. (6) is very similar to that of Eq. (3) except that a non-stationary process is now expressed as a time-varying summation of weighted complex exponentials. Meanwhile, the time-varying transfer function $H(t,f)$ can be replaced by the complex envelope or complex demodulate of process $Y(t)$ at the frequency of f. However, in practical situations $H(t,f)$ would be rather stochastic than deterministic due to the random temporal variations of the filter or the difficulty in acquiring precise time domain data. As such, a more comprehensive description of $Y(t)$ can be revised as

$$Y(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{j2\pi ft} H(t,f;\varpi) d\widehat{X}(f) \tag{7}$$

where $H(t,f;\varpi)$ denotes a complex envelope whose shape depends on the random variable ϖ. For simplicity, we impose that $H(t,f;\varpi)$ is time stationary and independent of the spectral process $d\widehat{X}(f)$. As a result, the fault signal $Y(t)$ we want to recover is categorized as processes of conditional non-stationarity (CNS). A few typical examples are regarded as the uniformly amplitude-modulated processes

$$Y(t) = m(t;\varpi)[h(t) * X(t)], \tag{8}$$

where $h(t,s;\varpi) = m(t;\varpi)h(s)$, randomised cyclostationary processes

$$h(t,s) = h(t+T,s) = \sum_k h_k(s)\mathrm{e}^{j2\pi kt/T}, \tag{9}$$

where T is a given period, and generalized shot noise

$$h(t,t-\tau;\varpi) = h(t-\tau)\sum_k \delta(\tau - \tau_k(\varpi)), \tag{10}$$

where $X(\tau_k)$ determines the random amplitude of the pulses. More specific cases are available in Refs. [11, 13].

2.2 Spectral Kurtosis

In the cases of conditional non-stationary processes mentioned above, spectral kurtosis can be formally defined as the energy-normalized forth-order spectral cumulant

$$SK_Y(f) = \frac{S_{4Y}(f)}{S_{2Y}^2(f)} - 2, \quad f \neq 0, \tag{11}$$

where the 2n-order spectral moments is defined as

$$S_{2nY}(f) \triangleq E\{S_{2nY}(t,f)\} = E\left\{|H(t,f)dX(f)|^{2n}\right\} = E\left\{|H(t,f)|^{2n}\right\} \cdot S_{2nX}, \quad (12)$$

Considering the actual vibration measurement being $Z(t) = Y(t) + N(t)$, the SK can be expressed as

$$SK_Z(f) = \frac{SK_Y(f)}{[1+\rho(f)]^2} + \frac{\rho^2(f)SK_N}{[1+\rho(f)]^2}, \quad f \neq 0 \quad (13)$$

where $\rho(f) = S_{2N}(f)/S_{2Y}(f)$ is the noise-to-signal ratio. Specially, if $N(t)$ is an additive stationary Gaussian noise independent of $Y(t)$, the spectral kurtosis of $Z(t)$ is simplified as

$$SK_Z(f) = \frac{SK_Y(f)}{[1+\rho(f)]^2}, \quad f \neq 0 \quad (14)$$

2.3 Illustration Example of Spectral Kurtosis

To illustrate the effectiveness of spectral kurtosis, a synthetic signal $x(t)$ containing repetitive impulsive transient components and white Gaussian noise is used. The sampling length of the signal is 50,000, and the theoretical spectral content of the transients is located in the normalized frequency band of [0.15, 0.19]. The time domain waveform of $x(t)$ is shown in Fig. 1. The masked impulsive transients are difficult to be identified by visual inspection.

The power spectrum density (PSD) function of the original signal is displayed in Fig. 2, from which it is seen that there are no spectral sidebands that helps to identify the repetitive transients.

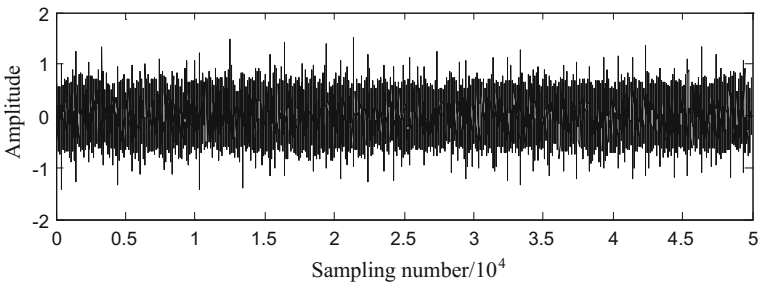


Fig. 1 Time domain waveform of the synthetic signal

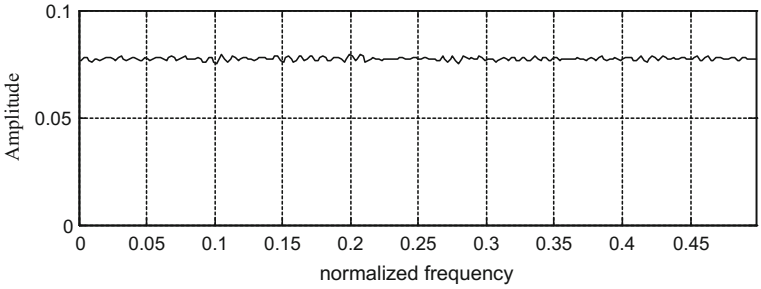


Fig. 2 Time power spectrum density of the synthetic signal

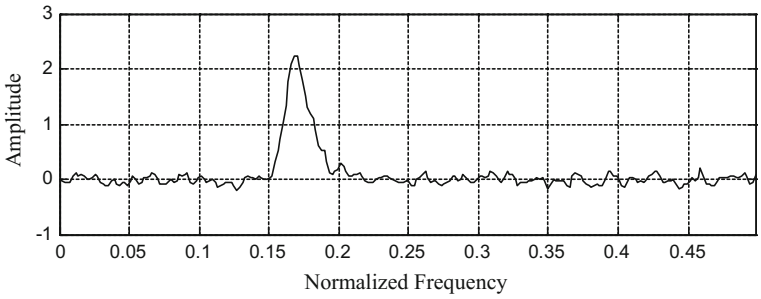


Fig. 3 Time spectral kurtosis distribution of the synthetic signal

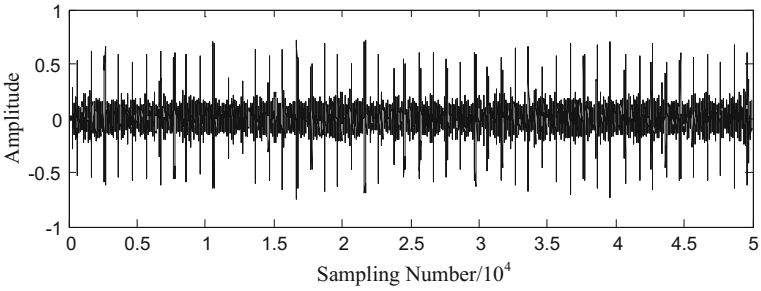


Fig. 4 Filtered signal containing repetitive impulsive transients

The spectral kurtosis distribution of $x(t)$ is computed and shown in Fig. 3. It is evident to recognize a global maximum point at the normalized frequency of 0.17. The spectral kurtosis distribution is used as the optimal filter and the corresponding filtered signal is shown in Fig. 4. It is observed that the transients of interest masked by white Gaussian noise are explicitly detected and extracted by the method of spectral kurtosis.

As comparison, we investigate the time domain kurtosis of the synthetic signal and the filtered signal. In statistics, the kurtosis indicates the fourth standardized moment, defined as

$$Kurt[X] = \frac{\mu_4}{\sigma^4} = \frac{E[(X-\mu)^4]}{(E[(x-\mu)^2])^2} \tag{15}$$

where μ_4 is the fourth moment about the mean and σ is the standard deviation. It is computed that $Kurt[x(t)] = 3.5022$, very close to the kurtosis of a statistically white Gaussian noise (3). While the kurtosis of the filtered signal in Fig. 4 is 11.0308. This comparison shows that usage and effectiveness of spectral kurtosis.

3 Wavelet Based Kurtogram and Its Development

Although SK is powerful in detecting impulsive transients in the presence of white Gaussian noise, its effectiveness can be very limited in processing vibration measurement $Z'(t)$ when interfering vibration modes are incorporated.

$$Z'(t) = Y(t) + N(t) + V(t) \tag{16}$$

where $V(t)$ indicates other vibration components independent of $Y(t)$. As can be inferred from Fig. 3, additional vibration modes would introduce other pass bands of the optimal filter, which result in mode mixing. In order to address this problem, pre-filtering step utilized all kinds of signal decomposition tools (detection filter) is required. A feasible detection filter is assumed to posses excellent time-frequency localizability whose properties are concluded in Lemma 1 [11].

Lemma 1 ***Requirement 1***. *Translation invariance: The values of the kurtogram are constraint to remain invariant to any time variation, i.e.* $\mathrm{SK}[x(n-n_0)] = \mathrm{SK}[x(n)]$.

Requirement 2. *Insensitivity to harmonics (discrete tones): The kurtogram of a discrete tone at frequency f0 is constraint to be nil (strictly speaking a discrete tone is stationary and therefore should not be detected by the kurtogram).*

Requirement 3. *Frequency localization: For the kurtogram to be interpreted as the kurtosis of the signal at a (frequency/frequency resolution) dyad* $\mathrm{AP}\{f, \Delta f\}$,[1] it is necessary that its estimator acts like a band-pass filter $[f - \Delta f, f + \Delta f]$.

Requirement 4. *Frequency concentration: For the kurtogram to be used to select the optimal frequency band where to demodulate a signal, it is necessary that its estimator fulfils that* $\Delta f \leq f$.

[12] AP means 'Analyzing Parameter'.

3.1 STFT Based Kurtogram

The initial detection filter used to estimate SK is STFT. For a non-stationary process $Y(t)$ with an analysis window $w(n)$ of length N_w and a give temporal step size P, the STFW is written as

$$Y_w(kP,f) = \sum_{n=-\infty}^{\infty} Y(n)w(n-kP)\mathrm{e}^{-j2\pi nf} \tag{17}$$

The 2n-order spectral moment of $Y_w(kP,f)$ is defined as

$$\hat{S}_{2nY}(f) = \left\langle |Y_w(kP,f)|^{2n} \right\rangle_k \tag{18}$$

with $\langle\cdot\rangle_k$ standing for the time-average operator over index k. Similar to Eq. (11)

$$SK_{STFT,Y}(f) = \frac{\hat{S}_{4Y}(f)}{\hat{S}_{2Y}^2(f)} - 2, |f - \mathrm{mod}(1/2)| > \frac{1}{N_W} \tag{19}$$

where N_w indicates the windows length. In STFT-based spectral kurtosis analysis (shown in Fig. 5), the STFT is used as a filterbank that involves frequency-shift, low-pass filtering, down-sampling operations, and the squared magnitude envelope of the filtered signal is used to estimate the peakiness of the transients at different frequency.

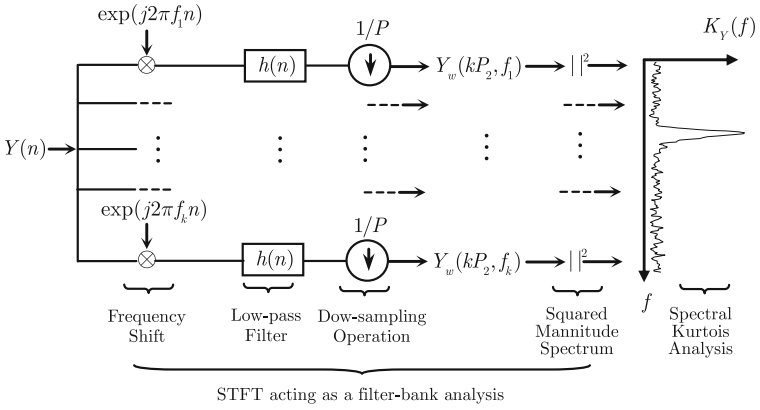


Fig. 5 Implementation of STFT-based spectral kurtosis

3.2 Fast Kurtogram

The four fundamental requirements in Lemma 1 are generally violated by DWT and WPT, a type of multi-rate filter-bank (MRFB) as the detection filter of the FK. The MRFB is composed of complex-valued finite impulse response (FIR) filters and approximates the four feasible properties. The implementing filter-bank of the MRFB is shown in Fig. 6, where the filter $\{h(n)|n = 0, 1, \ldots, L-1\}$ is a symmetric and low-pass filter. The complex-valued filters $\{h_0(n)\}$ and $\{h_1(n)\}$ are related as

$$h_1(n) = (-1)^{1-n} h_0(1-n) \tag{20}$$

A very notable advantage of FK over classical DWT and WPT is its ability to provide 1/3-binary tree decomposition to the signal, offering embedded frequency subbands of dyadic frequency partition grids.

In level k, the input signal is decomposed into 2^k distinct subbands. For the specific subband c_k^i, its Fourier spectrum is mainly localized in the frequency range of $[(i-1)\cdot 2^{-k}, i\cdot 2^{-k}]\pi$ and its central frequency is approximately equal to $(i-0.5)\cdot 2^{-k}\pi$. The frequency-scale paving plane of the original fast kurtogram is shown in Fig. 7.

3.3 Wavelet Packet Based Kurtogram

According to Ref. [15], the wavelet packet based kurtogram (WPBT) employs wavelet packet transform (the 'DB10' orthonormal basis, shown in figure) as its detection filter. Because the decomposition signals of WPT are real-valued, kurtosis, the traditional statistical indicator, is chosen to measure the impulsiveness of the signals rather than SK, even though the latter is also applicable to real-valued signals. Lei's filterbank inherits classical dyad-tree structure with DB10 wavelet

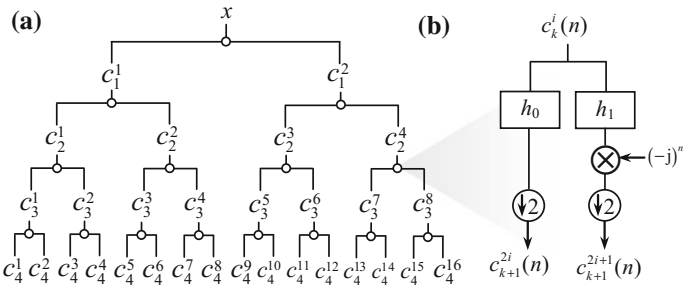


Fig. 6 Detection filter of the FK: **a** The MRFB structure of FK; and **b** function node of the multi-rate filter-bank

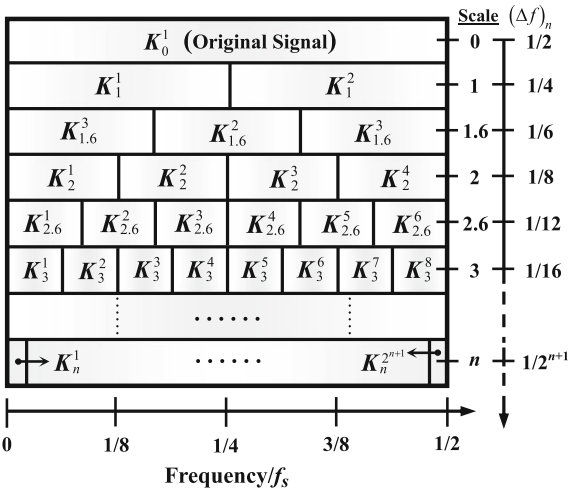


Fig. 7 The frequency-scale paving of the original fast kurtogram

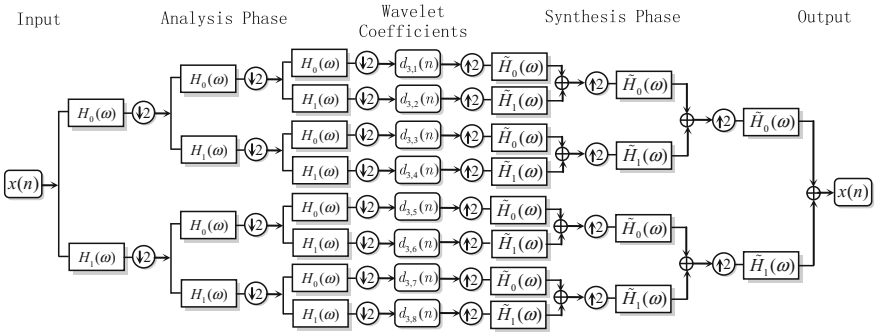


Fig. 8 The wavelet packet filterbank of WPBT

(Fig. 8). Moreover, Lei shown two applications and shown that WPBT demonstrated robuster performance than FK.

4 Wavelet Tight Frame Based Kurtogram

4.1 Limitation of Original Kurtogram

Although the original fast kurtogram is deliberately designed to possess the powerful vibration transients detecting ability characterized by the four properties listed in Lemma 1. However, in practical applications, three major limitations significantly hamper the effectiveness of the original fast kurtogram.

Limitation 1. Poor spatial resolution in higher decomposition stages

Owing to the down-sampling operations in the MRFB of the original fast kurtogram and its lack of perfect reconstruction property, the length of the series in higher decomposition stages are substantially truncated compared with that of the input signal. More importantly, the sampling density of the decomposition subbands becomes sparser, i.e. the spatial analyzing resolution decreases significantly, posing serious threats to successful investigations of transient signatures in these subbands.

Limitation 2. Relatively weak improvements in translation invariance

The MRFB of the FK was originally devised to possess better translation-invariance property. However, this improvement is not perfect. It turns out that the MRFB is far away form precise translation-invariance.

Limitation 3. Sensitivity of SK indicator to sporadic impulse

Despite the effectiveness of the SK indicator in detecting non-Gaussian transients, it is also sensitive to randomly occurred background noise and sporadic vibration components which are also be impulsive in nature. The sensitivity of the SK indicator to sporadic impulsive shocks will be shown via numerical trials and engineering applications.

4.2 Quasi-Analytic Wavelet Tight Frame

WPT offers finer frequency resolution in comparison with DWT, but inherits all mentioned limitations of DWT, some of which become even more severe in WPT. In this subsection we will adopt the strategy based on dual tree complex wavelet bases (Fig. 9) and construct a QAWPB combining the advantages of classical orthonormal basis and that of the orthogonal wavelet bases.

The hybrid filterbank for a QAWPT is shown in Fig. 10. Similar to that of DTCWT, the input signal is processed by two independent filter trees simultaneously. The filter-bank can be divided into three parts:

(1) The part of the first stage decomposition (FSD):

$$\{h_{10}^{\Re\mathrm{e}}(n), h_{11}^{\Re\mathrm{e}}(n), h_{10}^{\Im\mathrm{m}}(n), h_{11}^{\Im\mathrm{m}}(n)\};$$

(2) The part of dual three complex wavelet transform:

$$\{h_{0}^{\Re\mathrm{e}}(n), h_{1}^{\Re\mathrm{e}}(n), h_{0}^{\Im\mathrm{m}}(n), h_{1}^{\Im\mathrm{m}}(n)\};$$

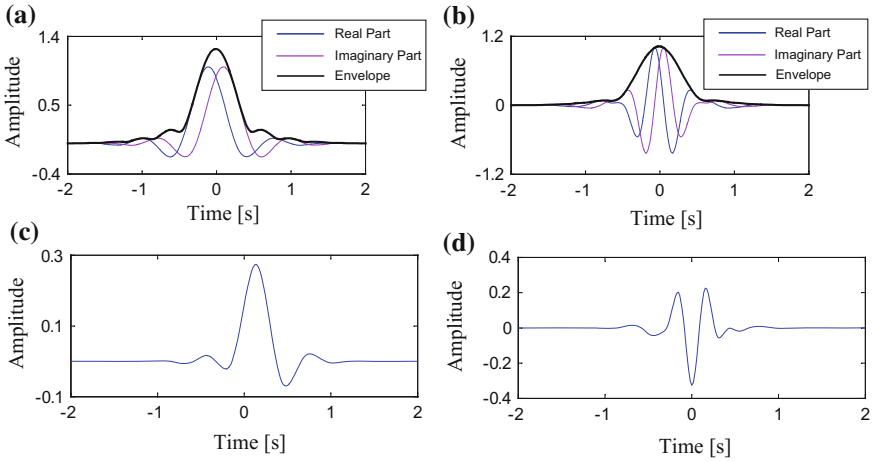


Fig. 9 Time-frequency atoms for the constructed DTCWT basis: **a** The complex scaling function and its envelope, **b** the complex wavelet function and its envelope; and time-frequency atoms of the 'Sym10' orthonormal wavelet basis, **c** the scaling function, and **d** the wavelet function

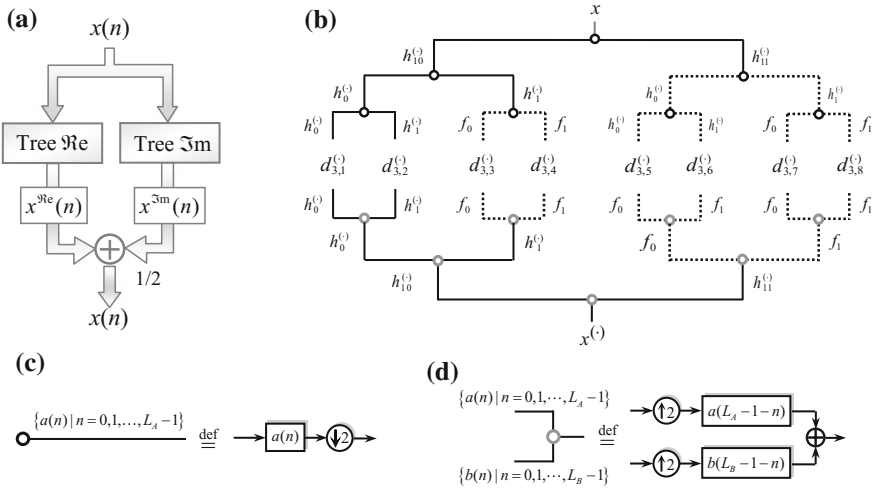


Fig. 10 Filter-bank for QAWPT **a** data flow of QAWPT; **b** implementing filter-bank for each filter branch (the superscript could be substituted by 'ℜe' or 'ℑm'); **c** function node in the analysis phase; and **d** function node in the synthesis phase

(3) The part of the generalized quasi-analytic wavelet packet (QAWP):

$$\{f_0(n), f_1(n)\}.$$

Different filter sets are allowed to be used for each individual part listed above; and theoretically speaking the filters can be chosen arbitrarily on condition that perfect reconstruction condition is satisfied. However, for acquiring better filtering performance, filters should be selected properly. Aiming at ensuring approximate linear-phase and obtaining reduced frequency aliasing, the scaling filter and wavelet filter originated from the 'sym10' orthonormal basis are also adopted for the QAWP part, i.e. the filters of non-DTCWT parts in Fig. 10 are defined as

$$\begin{cases} f_0(n) = h_{10}^{\Re\mathrm{e}}(n) \\ f_1(n) = h_{10}^{\Im\mathrm{m}}(n) \end{cases} \tag{21}$$

Let $W_{k,i}^{(\cdot)}(z)$, with k and i being positive integers, denotes the Z transform of the wavelet function corresponding to the node $d_{k,i}^{(\cdot)}$ and $n_{k-1}n_{k-2}\ldots n_1 n_0$ be the binary coding of the decimal integer $i-1$ such that

$$i - 1 = \sum_{m=0}^{k-1} 2^m n_m \tag{22}$$

where $n_m \in \{0, 1\}$ for $0 \le m \le k-1$. Then $W_{k,i}^{(\cdot)}(\mathrm{e}^{\mathrm{j}\omega})$, for $k \ge 1$, can be expressed as below:

$$W_{k,i}^{(\cdot)}(\mathrm{e}^{\mathrm{j}\omega}) = \begin{cases} H_{1,n_{k-1}}^{(\cdot)}(\mathrm{e}^{\mathrm{j}\omega}) \prod_{u=0}^{k-2} H_{nu}^{(\cdot)}(\mathrm{e}^{\mathrm{j}2^{k-1-u}\omega}) & i \in \{1,2\} \\ H_{1,n_{k-1}}^{(\cdot)}(\mathrm{e}^{\mathrm{j}\omega}) \left[\prod_{u=0}^{N_{\mathrm{one}}-1} F_{n_{k-1-u}}^{(\cdot)}(\mathrm{e}^{\mathrm{j}2^{k-1-u}\omega}) \right] \left[\prod_{v=N_{\mathrm{one}}}^{k-2} H_{n_{k-1-v}}(\mathrm{e}^{\mathrm{j}2^{k-1-v}\omega}) \right] & i \in \{3,4,\ldots,2^k\} \end{cases} \tag{23}$$

where N_{one} is the index of the first non-zero digit in the ordered sequence $\{n_{k-1}, n_{k-2}, \ldots, n_1, n_0\}$, i.e. $n_m = 0$ for $N_{\mathrm{one}} + 1 \le m \le k-1$ and $n_{N_{\mathrm{one}}} = 1$. Owing to the disadvantages brought about by the length truncation of the wavelet coefficient series, their corresponding single branch reconstruction signals are used for vibration analysis instead. After performing QAWPT on the input signal and single branch reconstruction on the complex wavelet coefficient series $d_{k,i}^{\mathbb{C}}(n) = d_{k,i}^{\Re\mathrm{e}}(n) + \mathrm{j} \cdot d_{k,i}^{\Im\mathrm{m}}(n)$, where $0 \le i \le 2^k - 1$, we can obtain the wavelet packet set $\{D_k^i(n) | i = 1, 2, \ldots, 2^k\}$. Let $W_{k,i}^{\Re\mathrm{e}}(\mathrm{e}^{\mathrm{j}\omega})$ and $W_{k,i}^{\Im\mathrm{m}}(\mathrm{e}^{\mathrm{j}\omega})$ denote the Fourier coefficient for the wavelet atoms corresponding to the wavelet packet $D_k^i(n)$. Reference [22] shows that

$$W_{k,i}^{\Im \mathrm{m}}(\mathrm{e}^{\mathrm{j}\omega}) \approx \mathrm{e}^{-\mathrm{j}0.5\omega} W_{k,i}^{\Re \mathrm{e}}(\mathrm{e}^{\mathrm{j}\omega}) \tag{24}$$

holds true for i = 1, 2. While for $i \in \{3, 4, \ldots, 2^k\}$, there exists that

$$\frac{W_{k,i}^{\Im \mathrm{m}}(\mathrm{e}^{\mathrm{j}\omega})}{W_{k,i}^{\Re \mathrm{e}}(\mathrm{e}^{\mathrm{j}\omega})} = \frac{W_{k-N_{\mathrm{one}},n_{N_{\mathrm{one}}}}^{\Im \mathrm{m}}(\mathrm{e}^{\mathrm{j}\omega})}{W_{k-N_{\mathrm{one}},n_{N_{\mathrm{one}}}}^{\Re \mathrm{e}}(\mathrm{e}^{\mathrm{j}\omega})} \tag{25}$$

where the subscript index $n_{N_{\mathrm{one}}}$ is ranged in {0, 1}. This leads to that

$$W_{k-N_{\mathrm{one}},n_{N_{\mathrm{one}}}}^{\Im \mathrm{m}}(\mathrm{e}^{\mathrm{j}\omega}) \approx \mathrm{e}^{-\mathrm{j}0.5\omega} \cdot W_{k-N_{\mathrm{one}},n_{N_{\mathrm{one}}}}^{\Re \mathrm{e}}(\mathrm{e}^{\mathrm{j}\omega}) \tag{26}$$

holds true for arbitrary subscript indices k and i, i.e. the resultant complex wavelet packet basis is also quasi-analytic. What's more, the QAWPT inherits the nearly translation-invariance of the DTCWT and has approximate linear-phase property.

Besides the convenience of automatic selection of optimal analyzing parameters, another significant advantage of the FK lies in that the MRFB is able to provide non-dyadic frequency subbands compared with DWT and WPT, therefore more effective in extracting fault features located in transition-areas of dyadic frequency partition grids. In order to achieve comparable feature extracting ability, we propose an ensemble wavelet subband generating strategy (EWSGS) that engenders non-dyadic wavelet subbands. The proposed EWSGS is based on the translation-invariance, perfect reconstruction and nearly linear-phase properties of QAWPT. The procedure of the proposed EWSGS is described as the following:

Step (i). Performing wavelet decomposition on the input vibration signal using the QAWPT and subsequently obtain the reconstructed wavelet packet set $\{D_k^i | i = 1, 2, \ldots, 2^k\}$.

Step (ii). Reordering the reconstructed wavelet packet set $\{D_k^i | i = 1, 2, \ldots, 2^k\}$ so that they are arranged in the ascending order with respect to their pass-band, i.e. $\{R_k^i | i = 1, 2, \ldots, 2^k\}$. We will show the mapping between these two sets as below: For a specific reordered wavelet packet $R_k^i(n)$ with the binary coding of $i - 1$ being

$$i - 1 = \sum_{m=0}^{k-1} 2^m n_m \tag{27}$$

define a decimal variable i' which can be expressed as

$$i' = 1 + \sum_{m=0}^{k-1} 2^m n'_m \tag{28}$$

where

$$n'_m = \begin{matrix} n_m & m = k-1 \\ \mathrm{mod}(n_m + n_{m+1}, 2) & m = 0, 1, \ldots, k-2 \end{matrix} \tag{29}$$

Then the series R_k^i in the reordered set is associated with the series $D_k^{i'}$ in the naturally reordered set of reconstructed wavelet packet signals.

Step (iii). Generating ensemble wavelet subbands using the following formula

$$ER_k^i(n) = R_k^{2i}(n) + R_k^{2i+1}(n), \quad \text{for } 1 \le i \le 2^{k-1} - 1 \tag{30}$$

4.3 Spatial-Spectral Ensemble Kurtosis and Its Kurtogram

Both of the kurtosis and the SK are less effective in distinguishing periodic impulses from other interfering vibration contents, especially when interfering contents such as background noise and sporadic vibrations also exhibiting impulsive natures in the time domain. However, their Fourier magnitude spectra stay irregular and produce low kurtosis values. This observation is not only effective in identifying Gaussian noise, but also in suppressing sporadic impulsive vibrations, which are usually broad-band distributed in the frequency domain (Figs. 11 and 12).

The proposed SSEK indicator attempts to make use of the information implied in the Fourier spectrum of the vibration signal to eliminate some interferences which cannot be correctly recognized by kurtosis and SK. Inspiringly, the SSEK indicator

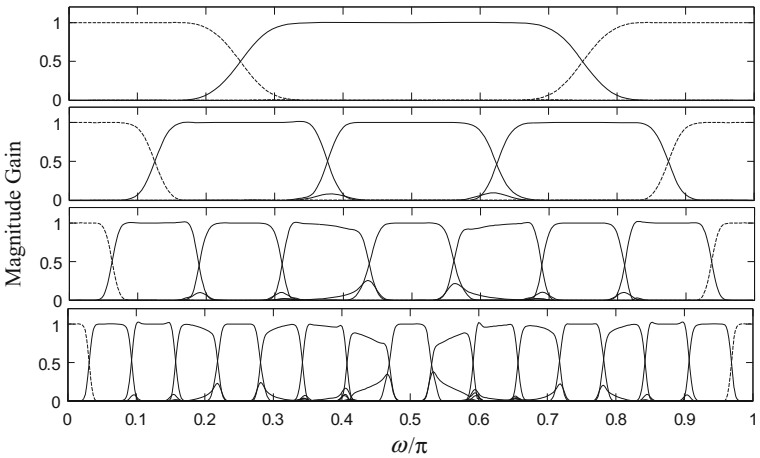


Fig. 11 Frequency response of the ensemble wavelet subbands (EWSs) derived form the first 5-stage quasi-analytic wavelet packet decomposition (the *purple cures* represent the frequency responses for the generated EWSs)

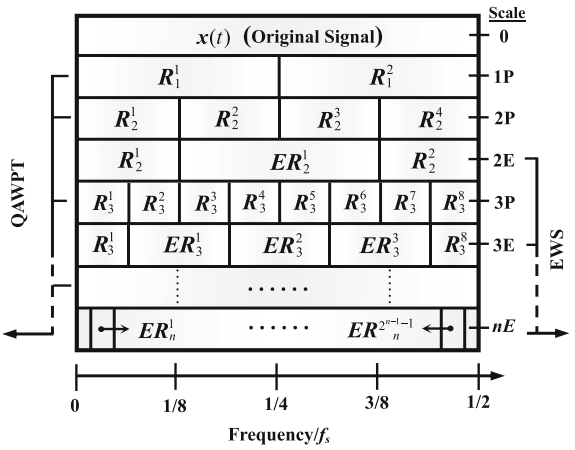


Fig. 12 The frequency-scale paving of the proposed SSEK kurtogram

is developed as a compound measure for evaluating the impulsiveness of vibration signals. For a specific wavelet subband, the SSEK value is also dependent on its frequency resolution. The SSEK indicator not only evaluates the impulsiveness of vibration signal in terms of its time-domain kurtosis, but also takes the kurtosis of its Fourier spectrum into consideration. Let $x(n)$ be a vibration signal of length L. The reordered wavelet packet set $\{R_k^i(n) \mid i = 1, \ldots, 2^k\}$ and the EWS set $\{ER_k^i(n) | \ i = 1, 2, \ldots, 2^{k-1} - 1\}$ are obtained by performing QAWTP and EWSGS on the input signal in sequence. The SSEK value for a specific wave packet R_k^i is defined as the following

$$\mathrm{SSEK}[R_k^i] \triangleq \mathrm{WFSK}[R_k^i] \cdot \frac{\sum_{n=\lfloor 0.05L \rfloor}^{\lfloor 0.95L \rfloor} \left(R_k^i(n) - \mu_{R_k^i} \right)^4}{(N-1)\sigma_{R_k^i}^4} \tag{31}$$

For a specific wavelet subband, its SSEK value is actually the product of its weighted Fourier spectrum kurtosis (WFMSK) index and its time-domain kurtosis. It can be checked that $\mathrm{SSEK}\left[D_{kw}^i\right] \geq 0$. Two additional specifications should be emphasized:

Remark 1 In computation of SSEK, not all coefficients of the reconstructed subband signal D_{kw}^i are used for calculating the time-domain kurtosis. Only the coefficients indexed within the interval $[\lfloor 0.05L \rfloor, \lfloor 0.95L \rfloor]$ are used so as to eliminate the inevitable boundary effects inherited in single branch wavelet reconstruction.

Remark 2 The purpose of the correction term is to endow the developed SSEK indicator with more robust identification ability of Gaussian noise, harmonics and sporadic impulses. For a particular wavelet subband D_{kw}^i, the WFMSK index is defined as the kurtosis the Fourier coefficients in set of

$$\{\hat{R}_k^i(f)|f \in [(i-1)\cdot k/2^{k+1}, i\cdot k/2^{k+1}]f_s\} \tag{32}$$

where $[(i-1)/2^{k+1}, i/2^{k+1}]f_s$ is the theoretical pass-band of the wavelet packet $R_k^i(n)$ in the frequency domain and f_s denotes the sampling frequency. The mathematical definition of the operator WFMSK[·] is given as

$$\mathrm{WFMSK}\left[R_k^i\right] \triangleq u\left(\mathrm{FMSK}\left[\hat{R}_k^i\right] - T_{low}\right)\cdot u\left(T_{upp} - \mathrm{FMSK}\left[\hat{R}_k^i\right]\right) \tag{33}$$

with

$$\mathrm{FMSK}\left[\hat{R}_k^i\right] \triangleq \frac{\int_{(i-1)/2^{k+1}f}^{i/2^{k+1}f}\left(\left|\hat{R}_k^i(f)\right| - \mu[\hat{R}_k^i]\right)^4 \mathrm{d}f}{(N-1)\sigma^4[\hat{R}_k^i]} \tag{34}$$

and $u(t)$ being the unit-step function, where $\mu[\widehat{R}_k^i]$ and $\sigma[\widehat{R}_k^i]$ are the mean and standard deviation of Fourier magnitude spectrum $\widehat{R}_k^i(f)$ within the theoretical pass-band of the input wavelet packet, i.e. $[(i-1)/2^{k+1}, i/2^{k+1}]f_s$; T_{low} and T_{upp} are threshold values. By definition, it can be deduced that WFMSK[·] is a bi-valued operator ranged in {0, 1}. For the derived EWS ER_k^i, its theoretical pass-band is assumed to be

$$[(2i-1)/2^{k+1}, (2i+1)/2^{k+1}]f_s \tag{35}$$

To compute the SSEK indicator, an important issue is the proper choice of T_{low} and T_{upp}. Based on a large amount of numerical simulations and abundant engineering experiences, the values of T_{low} and T_{upp} are determined to be 6 and 100 respectively Eq. (33) (Fig. 13).

4.4 Numerical Simulations and Engineering Applications

(a) ***Numerical simulations***

To verify their feasibility in the extracting transition-band features, a testing signal consisting of white Gaussian noise and periodically spaced impulses is simulated. The test signal is expressed as

$$x_{test}(t) = \sum_{i=1}^{10} L_i x_{ui}(t) + \mathrm{wgn}(t) \tag{36}$$

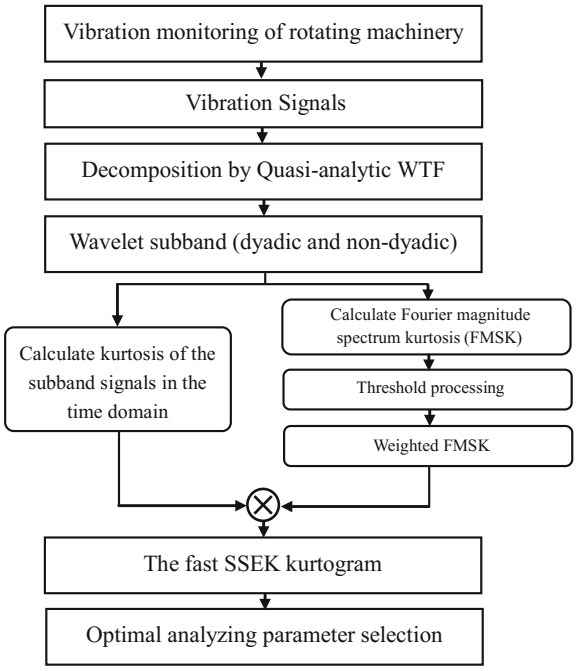


Fig. 13 Flow chart of the proposed SSEK kurtogram

where $x_{ui}(t) = \mathrm{e}^{-\beta(t-i\cdot \mathrm{T}_h)} \sin(2\pi 512t + \phi_i)$ for $1 \le i \le 10$; L_i denotes the amplitude of the ith impulse; $\mathrm{wgn}(t)$ is the white Gaussian noise series with certain intensity; the term $a_{ui}(t)$ represents the periodic amplitude modulation of the ith impulse; $\beta = 140$ is the damping characteristic of the system. The occurrence rate of the repetitive impulses $f_h = 1/T_h = 9.75$ Hz is set as non-integer because in engineering situations most occurrence rates of bearing faults and gear faults are fractional numbers. Moreover, the random variables $\{\phi_i | i = 1, 2, 3\}$, which are ranged in $(-\pi, \pi]$, are utilized to simulate the inconsistency inherent in the periodic impacts due to a variety of factors such as slip, varying load angle and transition path effect of engineering mechanical systems. The sampling rate of the testing signal is 2048 Hz and the sampling interval is 1 s. The original periodic impulses and its Fourier spectrum are shown in Fig. 14a, b, while the contaminated simulation signal containing white Gaussian noise and its Fourier spectrum are plotted in Fig. 14c, d. The signal-to-noise (SNR) ratio of the simulated signal is calculated to be −15.348 dB.

The proposed technique is applied to the analysis of the noise contaminated simulation signal, with the processing result shown in Fig. 15a. The optimal analyzing parameters are chosen as AP(512, 256 Hz). The corresponding optimal analysis subband is plotted in Fig. 15b, from which the masked periodic impulses are easily recognized. As comparison, the analyzing result by the WPT-based FK is shown in Fig. 15c and the optimal analyzing parameters are AP(480, 64 Hz), whose corresponding filtered signal shown in Fig. 15d reveals only a few scattered

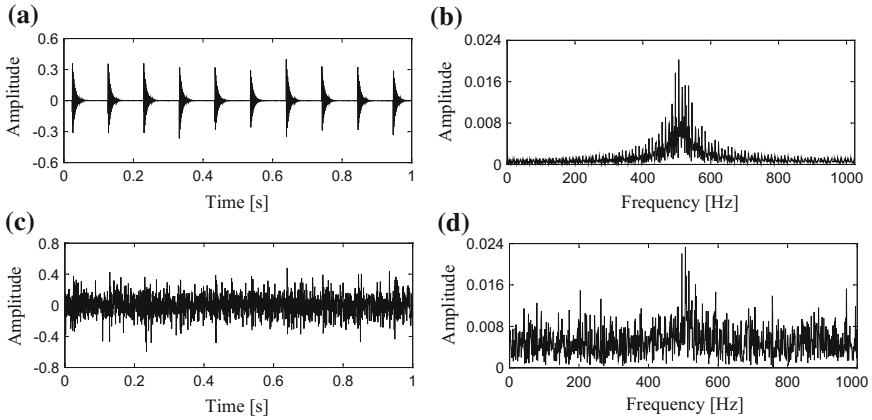


Fig. 14 **a** The periodic impulses; **b** Fourier spectrum of the periodic impulses; **c** the simulation signal containing periodic impulses masked by strong noise; and **d** the Fourier spectrum of the simulation signal in (**c**)

impulsive units. In contrasts, the processing result using the original FK is displayed in Fig. 15e. The filtered signal in Fig. 15f does not give so satisfactory result as that of the proposed technique. Although the harmonics existing in the envelope spectrum of the filtered signal indicates the existence of hidden transient vibration contents, the information in the time domain is not clear.

It is undeniable that each of the above three techniques detects the hidden fault features to a certain extent, which can be reflected in their envelop spectrum. However, the proposed technique is with the optimal analyzing result in the sense that it extracts the periodic impulses most successfully in the time domain, whereas the periodic impulses in the filtered signals other two contrasting methods are less clear.

(b) ***Engineering applications***

A three-axis NC horizontal boring and milling machine (HBMM), shown in Fig. 16a, was employed to conduct coaxial-hole boring of complex shell parts. The problem of low coaxiality of the machined holes was detected. By using another ordinary boring machine tool, the machining precision of the end face was examined again and the resultant coaxiality of the holes can meet the required machining precision. Therefore, it was inferred that the rotary table system of the HBMM was suffering major performance degradation. Because this NC machine tool had been put into service for more than ten years, the above phenomenon was considered to be natural. Ant the major issue was to locate the part suffering severe performance degradation and evaluate the remaining machining capacity of this HBMM.

The driving chain of the rotary table system is sketched in Fig. 16c. As illustrated, a servo motor is connected to the input shaft, and the torque is transmitted by the sequential combination of timing belts and worm gears. The rotary table is

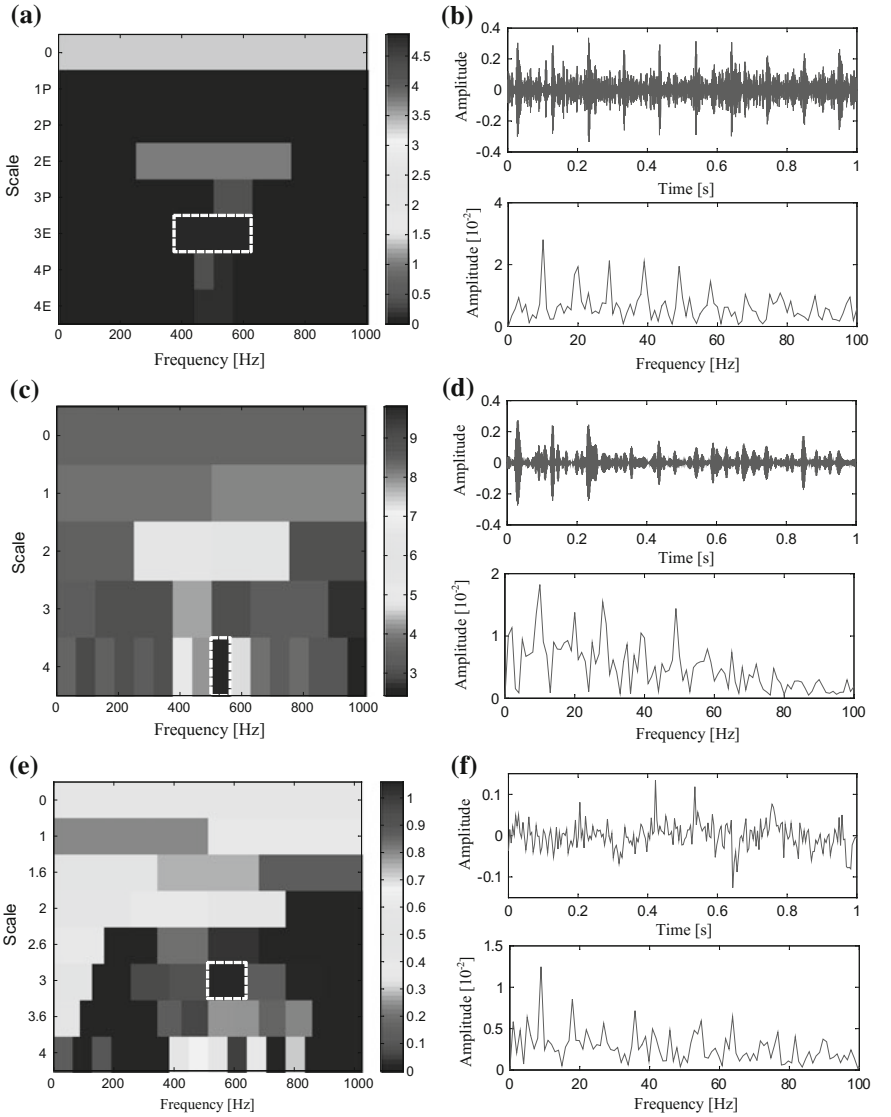


Fig. 15 **a** SSEK kurtogram of the simulation signal **b** the optimal analysis subband and its envelope spectrum of SSEK kurtogram **c** DWT based fast kurtogram of the simulation signal **d** the optimal analysis subband and its envelope spectrum of DWT based fast kurtogram **e** original fast kurtogram of the simulation signal and **f** the optimal analysis subband and its envelope spectrum of original fast kurtogram

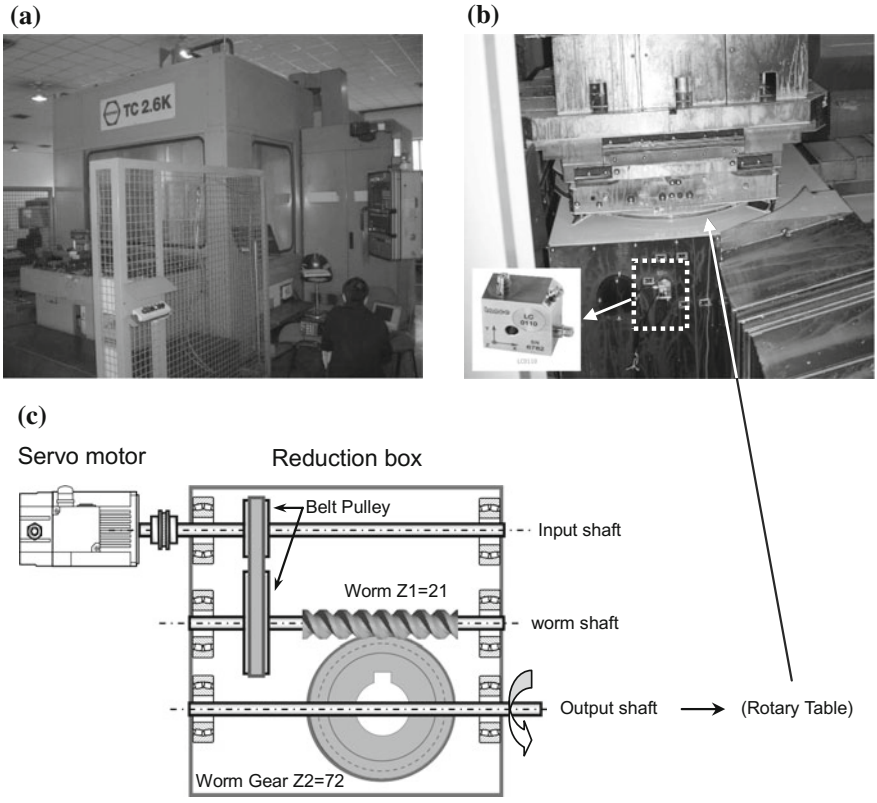


Fig. 16 **a** Photograph of the HBMM, **b** the mounted sensor, **c** the schematic diagram of the transmission chain

connected to the output shaft on which the worm gear is installed. A vibration measurement system was set up to collect the vibration signals during the HBMM's machining operation, as shown in Fig. 17b. A 3-axis accelerometer was mounted on the housing of the rotary table and the acceleration signals were collected. The acceleration signals were sampled at 12.8 kHz. Each record of collected signal is of length 16,384. When the rotary table was operating at constant speed, it was measured that the rotation speed of the servo motor was 2000 r/min. Accordingly, the expected characteristic frequencies of the rotary table system are listed in Table 2.

The acceleration signal of the X-axis vibration and its Fourier spectrum are shown in Fig. 17. As it can be seen, many interfering contents exist in time domain and frequency domain and mask the critical information related to the HBMM's

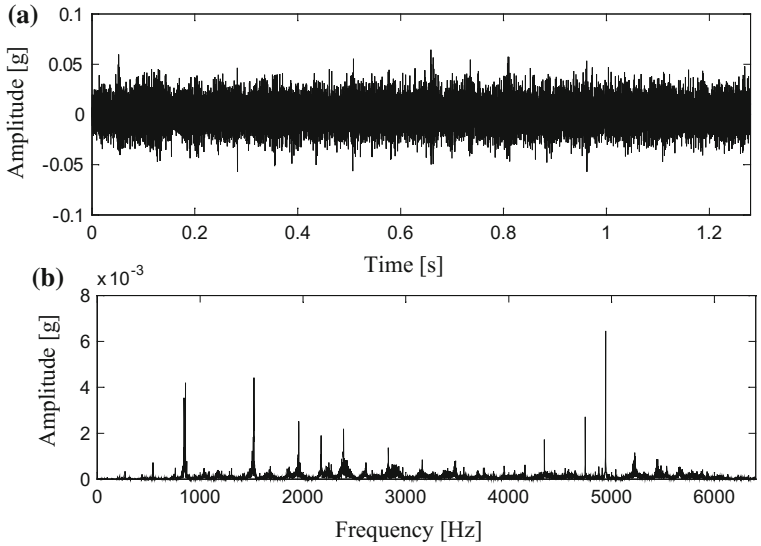


Fig. 17 **a** Vibration signal of rotary table and **b** Fourier spectrum of the rotary table vibration signal

Table 2 Expected characteristic parameters of the reduction gearbox in the milling machine

Transmission ratio [Z2/Z1]	65/22
Rotation frequency of input shaft [Hz]	4.5
Rotation frequency of output shaft [Hz]	1.52
Meshing frequency [Hz]	99
Module of the pinion [mm]	30
Central distance [mm]	1350
Face-width [mm]	560

Table 3 Expected characteristic parameters of the reduction gearbox of the HBMM

Rotation speed of the servo motor [Hz]	33.33
Transmission ratio of the timing belts	2.5
Rotation speed of the worm shaft [Hz]	13.33
Transmission ratio of worm gear pair	72
Rotation speed of the worm gear [Hz]	0.185

working condition. What's more, the frequency peaks of 857.80, 1523.00 Hz and etc. are not found in Table 3, hence no critical vibration signatures related to rotary table system is recognized.

The SSEK kurtogram of the acceleration signal is shown in Fig. 18a, in which the optimal analyzing parameters AP(3000, 400 Hz) are selected. The time domain waveform and the envelope spectrum corresponding to the optimal analyzing parameters are shown in Fig. 18b. It can be observed that there are repetitive impulses spaced at 0.075 s. This impulse occurrence rate is exactly the same as the

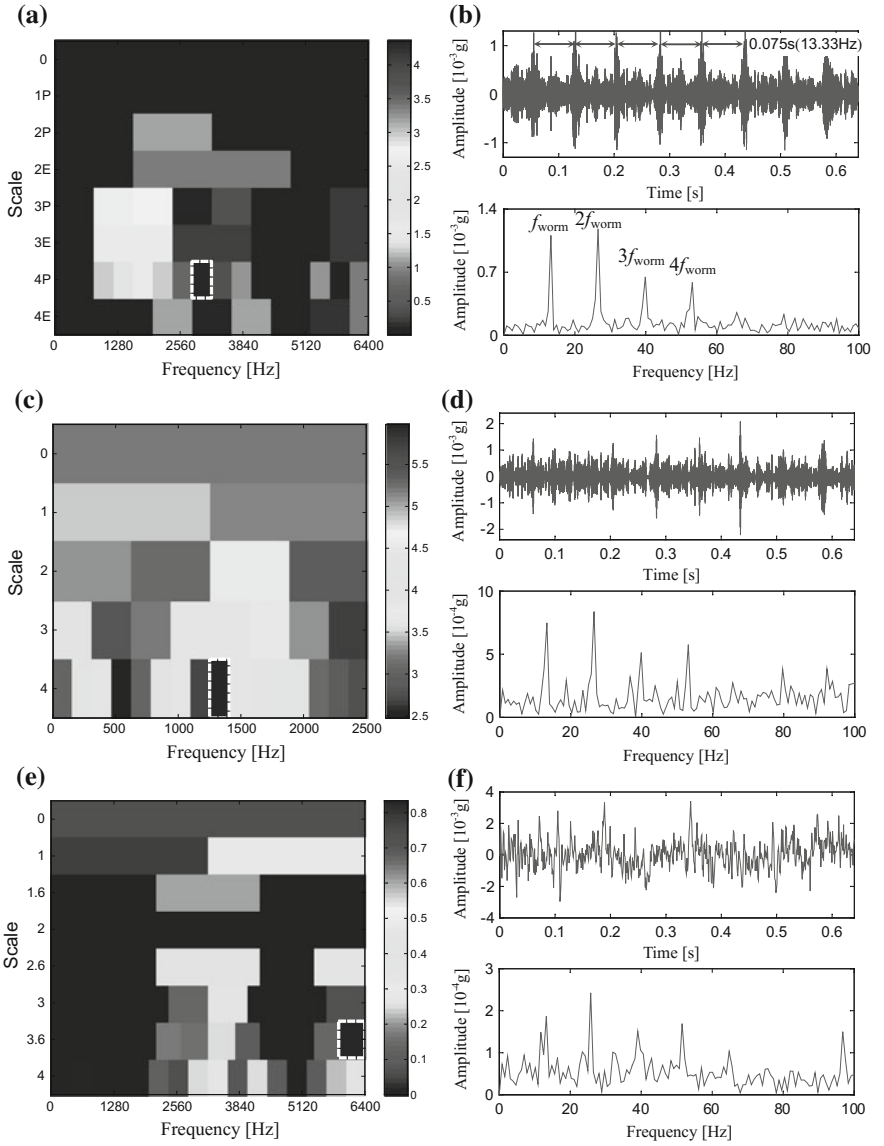


Fig. 18 **a** SSEK kurtogram of the rotary table's acceleration signal; **b** the optimal analysis subband and its envelope spectrum of SSEK kurtogram; **c** DWT based fast kurtogram of the rotary table's acceleration signal; **d** the optimal analysis subband and its envelope spectrum of DWT based fast kurtogram; **e** original fast kurtogram of the rotary table's acceleration signal; and **f** the optimal analysis subband and its envelope spectrum of original fast kurtogram

rotation frequency of the worm shaft, meaning that the impulses happened every single revolution of the worm. Thus it was surmised that the worm suffered localized fault on its surface, which caused abnormal vibration during the HBMM's machining process and resulted in low machining precision. This speculation was validated in an afterwards overhaul of the HBMM. Without available spare part of worm of this type, the HBMM, after long service term, was determined to be no longer suitable for high precision machining. However, it is still feasible for semi-finishing machining purposes.

The other two contrasting methods are also employed in this case study, with their processing results shown in Fig. 18c–f. It can be seen form Fig. 18d that the optimal analysis subband of the WPT-based FK only reveals a few isolated and irregular impulses. The periodicity of the impulses in Fig. 18d is not so evident as that shown in Fig. 18b. On the other hand, the FK selects the high frequency noise as its optimal analysis subband. Although the optimal analysis subband of the FK its envelope spectrum disclose there may be periodic phenomenon, its time domain waveform does not provide and explicit information about that.

5 Adaptive Super-Wavelet Based Kurtogram

Sparse representation of fault features is of great importance to the feature extraction of machinery fault detection [23]. Generally, the detection filter used for estimation of spectral kurtosis are based on fixed bases or frames. In this section, an adaptive detection filter name ensemble super-wavelet (ESW) is applied. The ESW is put forward based on the combination of tunable Q-factor wavelet transform (TQWT, [24]) and Hilbert transform such that fault feature extracting adaptability is enabled.

5.1 Adaptive Super-Wavelet Transform

(a) ***Filter bank***

The TQWT is implemented using perfect reconstruction over-sampled filter banks with real-valued sampling factors. The transform consists of a sequence of two-channel filter banks, with the low-pass output of each filter bank being used as the input to the successive filter bank. The associated analysis and synthesis filter banks for the TQWT are shown in Fig. 19 (where LPS α and HPS β represent low-pass scaling and high-pass scaling with the parameters α and β, respectively). The filter bank parameters α and β should be properly chosen so as to achieve a wavelet transform with the desired Q-factor and over-sampling rate r. Here, the relationship between (α, β) and (Q, r) can be expressed as

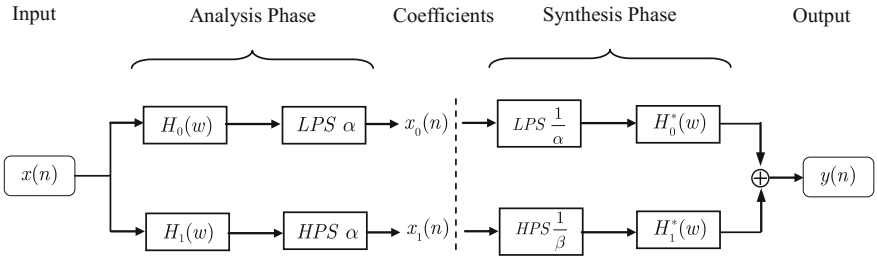


Fig. 19 Filterbank of TQWT

$$\beta = \frac{2}{Q+1}, \quad \alpha = 1 - \frac{\beta}{r}. \tag{37}$$

For the purpose of perfect reconstruction, the frequency responses $H_i(\omega), i = 0, 1$, must be chosen so that the reconstruction signal $y(n)$ equals the input signal $x(n)$. Specifically, the frequency responses of $H_0(\omega)$ and $H_1(\omega)$ are defined as

$$H_0(\omega) = \begin{cases} 1, & |\omega| \le (1-\beta)\pi \\ \theta(\frac{\omega+(\beta-1)\pi}{\alpha+\beta-1}), & (1-\beta)\pi \le |\omega| < \alpha\pi \\ 0, & \alpha\pi \le |\omega| \le \pi \end{cases} \tag{38}$$

$$H_1(\omega) = \begin{cases} 0, & |\omega| \le (1-\beta)\pi \\ \theta(\frac{\alpha\pi-\omega}{\alpha+\beta-1}), & (1-\beta)\pi \le |\omega| < \alpha\pi \\ 1, & \alpha\pi \le |\omega| \le \pi \end{cases} \tag{39}$$

where

$$\theta(\omega) = \frac{1}{2}(1+\cos\omega)(2-\cos\omega)^{1/2} \quad \text{for } |\omega| \le \pi \tag{40}$$

The transition function $\theta(\omega)$, originating from the Daubechies filter with two vanishing moments, is used to construct the transition bands of $H_0(\omega)$ and $H_1(\omega)$. It can be verified that the low-pass filter $H_0(\omega)$ and high-pass filter $H_1(\omega)$ satisfy the perfect reconstruction requirement $H_0(\omega)\}^2 + |H_1(\omega)|^2 = 1$. The variables α and β are the LPS parameter and the HPS parameter, respectively. They satisfy

$$\begin{cases} 0 < \alpha < 1 \\ 0 < \beta \le 1 \end{cases} \tag{41}$$

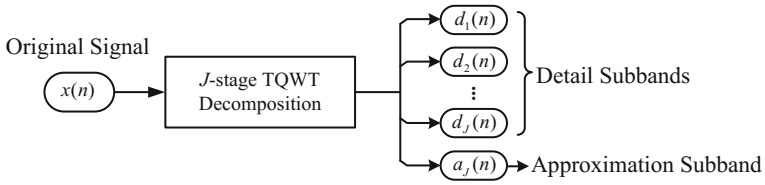


Fig. 20 TQWT with J-stage decomposition

so as to ensure the wavelet transform will not be overly redundant. In order that the filter responses be well localized, it is necessary that $\alpha + \beta > 1$.

(b) ***TQWT parameters***

The TQWT is specified by three parameters: the Quality-factor Q, the redundancy r, and the number of stages (or decomposition levels) J (Fig. 20).

Generally, Q is a measure of the number of oscillations the wavelet exhibits. According to the definition of the Q-factor, the bandwidth varies inversely to the Q-factor for a given center frequency. Therefore, a higher Q-factor has a better frequency resolution in comparison with a lower one. For Q, a value of 1.0 or greater can be specified. The wavelet with a Q-factor 4.0 or greater consists of sufficient oscillatory cycles to process signals with oscillatory features. Meanwhile, for Q, a value of 4.0 or greater is large enough for the precise division of frequency band of mechanical vibration signals. Therefore, the lower and upper bounds of the Q-factor are set to be 1.0 and 5.0 respectively in this article. The parameter r is the redundancy of the TQWT when it is computed using infinitely many levels. The specified value of r must be greater than 1, and a value of 3 or greater is generally recommended. The parameter J denotes the number of filter banks. As illustrated in Fig. 20, after performing a J-stage TQWT decomposition, there will be $J + 1$ subbands: the high-pass filter output signal of each filter bank $\{d_j(n)|j = 1, \ldots, J\}$, and the low-pass filter output signal of the final filter bank $a_J(n)$.

(c) ***TQWT wavelets and frequency responses***

Figure 21 shows the TQWT for two different sets of TQWT parameters. By increasing the value of Q, the wavelet function becomes more oscillatory, as illustrated in Fig. 21 Obviously, it can be observed that for $Q = 1.0$ the wavelet consists of few oscillatory cycles compared with the wavelet for $Q = 3.5$. Meanwhile, the frequency responses of the band-pass filters constituting the TQWT can be easily observed in Fig. 21. For a low Q-factor, the band-pass filters are quite wide. In contrast, for a high Q-factor, the filters are narrower. Therefore, a high Q-factor wavelet transform requires more levels to cover the same frequency range as a low Q-factor transform [25].

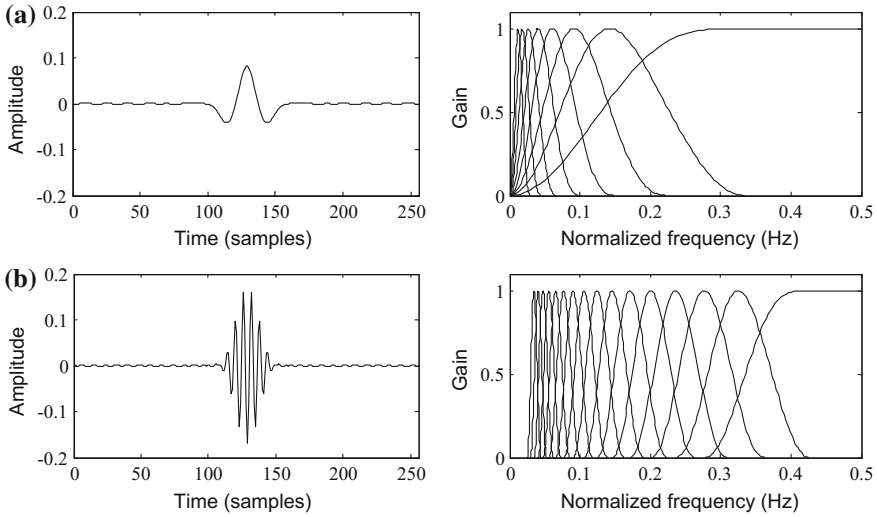


Fig. 21 TQWT wavelets and frequency responses (at level 6) for two Q-factors: **a** $Q = 1$, $r = 3$, $J = 8$; and **b** $Q = 3.5$, $r = 3$, $J = 16$

In addition, the TQWT can be implemented efficiently using radix-2 FFTs. The computational cost of the TQWT is as low as expected, given the implementation is based on the discrete Fourier transform.

5.2 A Sparse Indictor: Fault Feature Ratio (FFR)

The FFR measures the peakiness of a non-stationary signal in the Hilbert envelope domain [26]. A record of real signal is defined as the real part of an analytic signal

$$x^{\mathrm{Re}}(t) \stackrel{\Delta}{=} x(t) \tag{42}$$

while its counterpart in the Hilbert domain is expressed as

$$x^{\mathrm{Im}}(t) = \pi^{-1} \int_{-\infty}^{+\infty} x^{\mathrm{Re}}(\tau) \frac{1}{t-\tau} \mathrm{d}\tau \tag{43}$$

In consideration of the slow decaying property of convolution function $g(t) = 1/(\pi t)$ in the time domain, the above operation are often made in the Fourier domain.

$$\hat{x}^{\mathrm{Im}}(\omega) = \begin{cases} -\mathrm{j} \cdot \hat{x}^{\mathrm{Re}}(\omega), & \omega > 0 \\ \mathrm{j} \cdot \hat{x}^{\mathrm{Re}}(\omega), & \omega < 0 \end{cases} \tag{44}$$

As for complex-valued signal $x^{\mathbb{C}} = x^{\mathrm{Re}}(t) + \mathrm{j} \cdot x^{\mathrm{Im}}(t)$, the instantaneous amplitude $I_{Amp}(t)$ and the instantaneous freuqnecy $I_{Freq}(t)$ can be computed as

$$I_{Amp}(t) = \left[(x^{\mathrm{Re}}(t))^2 + (x^{\mathrm{Im}}(t))^2 \right]^{1/2} \tag{45}$$

$$I_{Freq}(t) = \arctan\left[x^{\mathrm{Im}}(t) / x^{\mathrm{Re}}(t) \right] \tag{46}$$

The Hilbert envelope spectrum $I_{Amp}(t)$ is the Fourier transform of $I_{Amp}(t)$

$$H\{I_{Amp}(t)\} = \mathrm{DFT}\{I_{Amp}(t)\} \tag{47}$$

For a specific characteristic frequency f_{Cha} ,the energy weight of its higher order harmonic $i \cdot f_{Cha}$ in the Hilbert envelope $H\{I_{Amp}(t)\}$ is computed as

$$E_{imp}(x, f_{Cha}) = \frac{2}{L} \cdot \frac{\sum_{i=1}^{i \cdot f_{Cha} \le f_s/2} \max_{k \in I_{iChar}} \left\{ \left| \sum_{n=0}^{L-1} x(n) \exp\left(-j \frac{2\pi}{L} nk\right) \right| \right\}}{\sum_{n=0}^{L/2-1} \left[\left| H\{I_{Amp}(t)\} \right| \right]} \tag{48}$$

where f_s denotes the sampling frequency of $x(t)$, and I_{iCha} denotes the index interval of the ith order harmonic component of f_{Cha}

$$I_{iCha} = \left\{ k | k = \frac{iLf_{Cha}}{f_s} \pm m,\ m = 0, \ldots, 10 \right\} \tag{49}$$

According to the characteristics of impulsive transients, we choose the subset of $i \in \{1, 2, 3\}$. Thus Eq. (49) can be simplified as

$$\tilde{E}_{imp}(x, f_{\boldsymbol{Cha}}) = \frac{|S(f_{\mathrm{Cha}})| + |S(\boldsymbol{2} f_{\mathrm{Cha}})| + |S(\boldsymbol{3} f_{\mathrm{Cha}})|}{\sum_{f=0}^{f_s/\boldsymbol{2}} |S(f)|} \tag{50}$$

where $S(f)$ denotes the amplitude of nth order harmonic component of f_{Cha}. Equation (50) calculates the overall energy of $x(t)$ as $f \to \infty$.

5.3 Adaptive ESW Based Kurtogram

The procedure of the ESW based kurtogram is illustrated in Fig. 22. The proposed method is composed of three major steps:

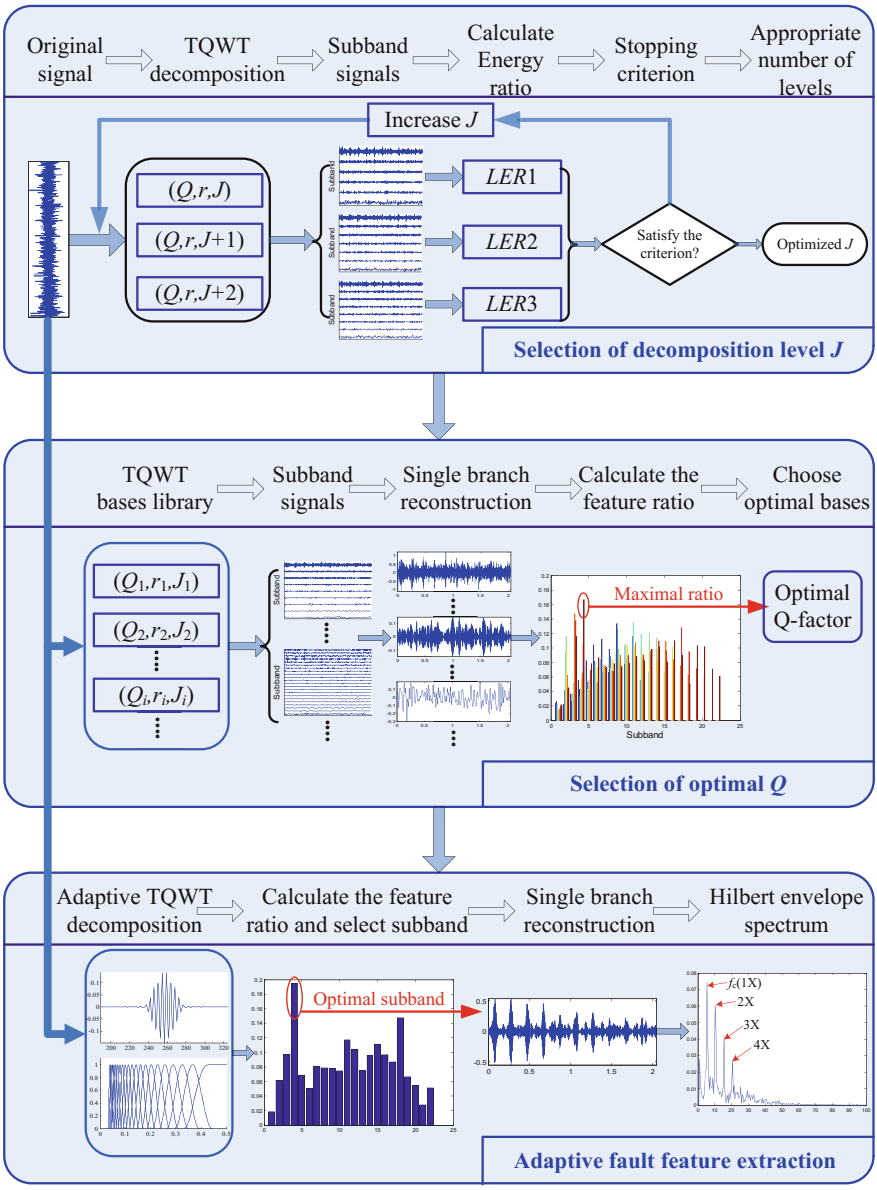


Fig. 22 The procedure of the proposed adaptive feature extraction method based on the TQWT

(1) Selection of decomposition level J;
(2) Selection of optimal Q-factor;
(3) Adaptive fault feature extraction.

Note that the value of r is specified to be 3.0. The detailed descriptions of these steps are presented in the following subsection.

(a) ***Selection of decomposition level J***

For a given TQWT parameter pair (Q, r), the maximum number of decomposition levels is

$$J_{\max} = \left| \frac{\log(\beta N/8)}{\log(1/\alpha)} \right| \tag{51}$$

where α and β are the low-pass scaling parameter and the high-pass scaling parameter, N is the length of the input signal. The maximum number of decomposition levels $J_{\max}$ increased with an increase in Q or N.

Although TQWT can be efficiently implemented using radix-2 FFTs, too much computational time introduced by a large number of decomposition levels is still computationally intractable, especially when different Q-factors are used to analyze large data. Meanwhile, excessive decomposition levels may result in inappropriate decomposition of useful signatures or redundant processing of useless signal components. Therefore, an appropriate decomposition level J is needed. Addressing this problem, a decomposition stopping criterion is devised to select the decomposition level J adaptively. The relationship between the energy ratio of last subband to the total energy (*LER*) and decomposition level J when applying TQWT to process dynamic signals is illustrated in Fig. 22. Here for the convenience of the comparative, the *LER* on the vertical axis remains equal to the value calculated when the decomposition level is $J_{\max}$ if J on the horizontal axis exceeds $J_{\max}$. The *LER* is given by

$$LER = \frac{E_{J+1}}{E_{tot}} \tag{52}$$

where E_{J+1} is the energy of the last low-pass subband (namely the $J + 1$ subbands) after performing a J-stage TQWT decomposition, E_{tot} is the total energy in wavelet domain. It can be clearly observed from Fig. 22 that the *LER* decreased as decomposition level J increased. The stopping criterion is based on the assumption that if a proper decomposition level J is obtained, then the *LER* keeps nearly unchanged. Here the relative difference of adjacent *LER* constrained is used to obtain a threshold T to achieve a proper decomposition level J. Specifically, an appropriate J is obtained when the following stopping criterion is satisfied.

$$|LER1 - LER2| + |LER1 - LER3| \leq T \tag{53}$$

where $LER1$, $LER2$ and $LER3$ are the energy ratio corresponding to J, $J + 1$ and $J + 2$ levels. The threshold value T should be set appropriately. Generally, T can be set as 2%. Using a smaller value will result in more decomposition levels. Using a greater value will result in less decomposition levels. Note that setting an appropriate decomposition level is always a difficult task when using wavelet transform to extract fault features in engineering applications.

To reduce the computation cost, the initial decomposition levels J should be set. In this case, the number of initial decomposition levels corresponding to $Q = 1.0$ is chosen to be 5, then one more level is needed if the Q is updated with an increase of 0.5.

The procedures of selecting an appropriate J can be summarized as follows.

Decompose the input vibration signal by using TQWT bases into $J + 1$, $J + 2$ and $J + 3$ (J represents the initial decomposition levels);
Calculate the energy ratio $LER1$, $LER2$ and $LER3$;

If the stopping criterion is not satisfied, then increase J and repeat steps (1) and (2) until an appropriate J is obtained to stop the circulation.

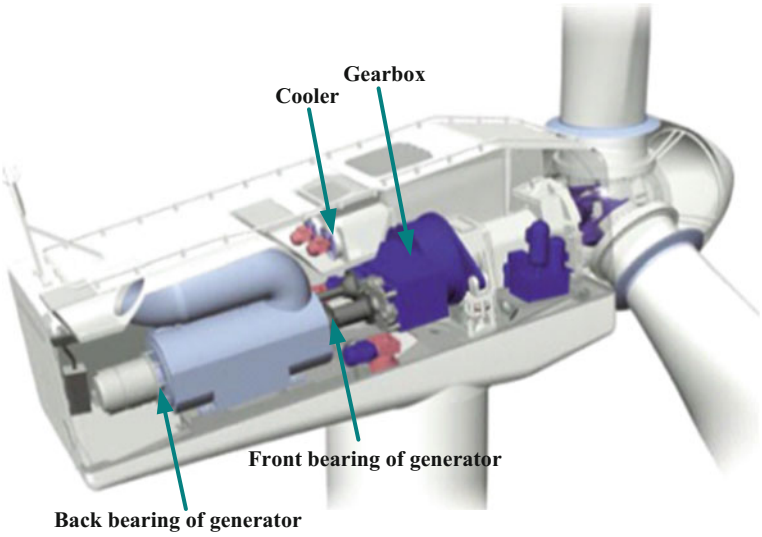


Fig. 23 Schematic draw a wind turbine system

Table 4 Parameters of the generator bearing in the wind turbine

Inner diameter (mm)	Outer diameter (mm)	Roller diameter (mm)	Number of rollers (mm)	Contact angle (mm)
120	280	41.275	8	0

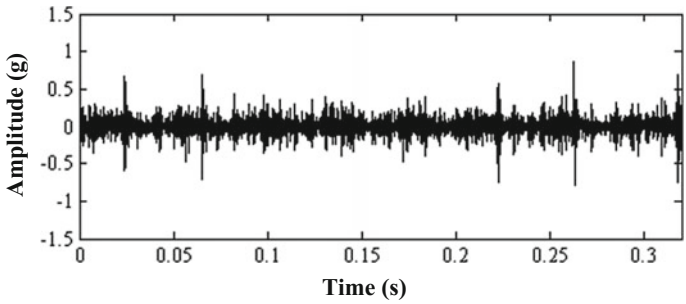


Fig. 24 The measured vibration signal

5.4 Engineering Applications

In this engineering application, a rolling element bearing fault developed in a wind turbine is detected by the proposed adaptive method. The structure sketch of a wind turbine is shown in Fig. 23, which mainly consists of a three-phase induction generator, a cooler, a single-stage planetary gearbox, etc. Both the front and back bearings of the generator are deep groove ball bearings. The bearing type is 6324 and its parameters are displayed in Table 4. Accelerometers for vibration detection were mounted on the housings of the front and back bearings. The sampling frequency is 12.8 kHz and the average rotating speed is 1501.2 r/min.

A measured vibration signal with a length of 4096 is shown in Fig. 24. However, only several aperiodic impulses can be revealed clearly in the time domain.

The frequency spectrum and the Hilbert envelope spectrum of the signal are shown in Fig. 25. The rotating frequency 25 Hz can be seen in Fig. 25a. In the envelope spectrum, the fault characteristic frequency of the front bearing inner race 125 Hz (calculated value $f_i = 120.74$ Hz) can be found, but it is surrounded by heavy noise frequencies and its harmonic components are not revealed. Therefore, the vibration signal and the spectral analysis cannot provide useful diagnosis information effectively.

The proposed method is introduced to analyze this vibration signal and extract the useful fault features hidden in the signal. Figures 26 and 27 show the processing results. Here the optimal Q-factor is chosen to be 2.5. It can be seen from Fig. 27a, strong periodic impulses with intervals of 0.04 s are clearly revealed, which is exactly in accordance with the rotating frequency 25 Hz. Moreover, we can find some weak periodic impulses, which is relevant to the fault characteristic frequency of inner race

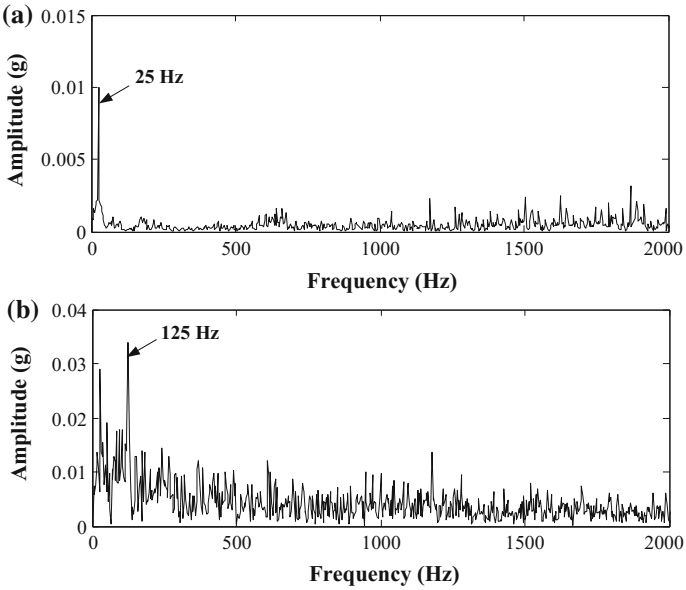


Fig. 25 **a** Fourier spectrum; and **b** Hilbert envelope spectrum

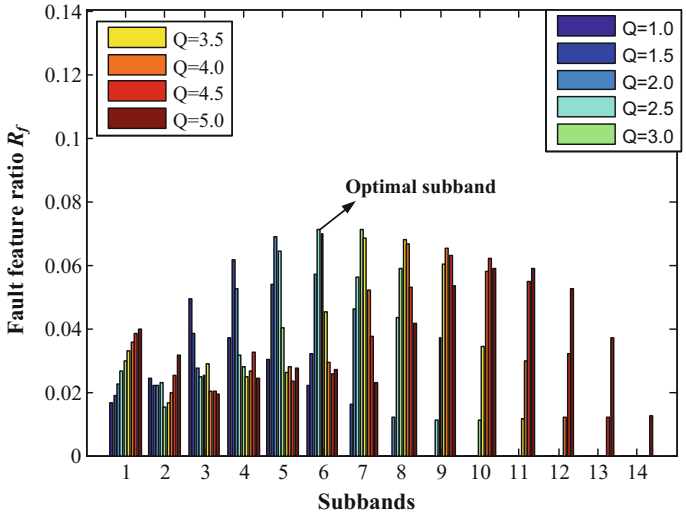


Fig. 26 Fault feature ratio R_f displayed along the subbands corresponding to different Q-factors in case 2

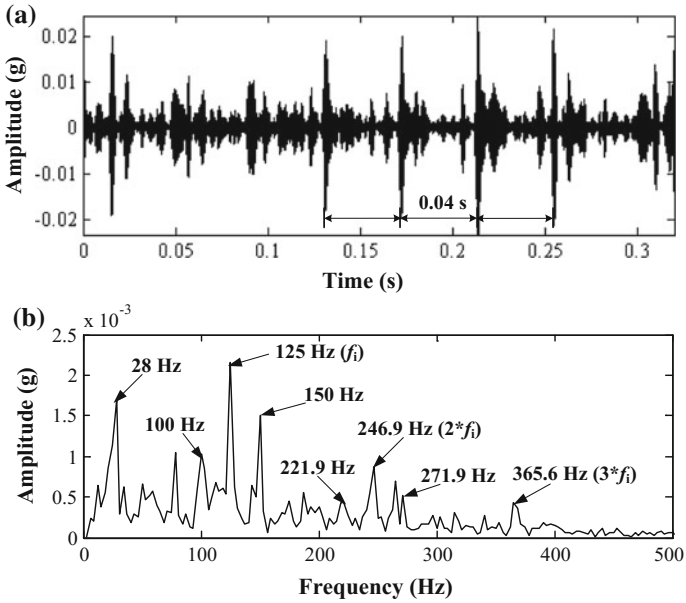


Fig. 27 Fault feature extraction results: **a** single branch reconstruction of the selected optimal subband; and **b** Hilbert spectrum of the reconstructed signal

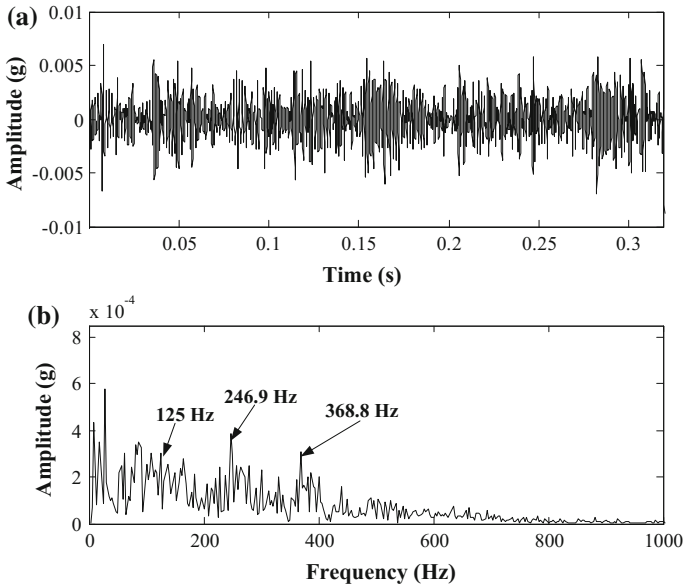


Fig. 28 The analyzed result in case 2 by using non-adaptive TQWT feature extraction method: **a** single branch reconstruction of the selected subband; and **b** Hilbert envelope spectrum

f_i. As shown in Fig. 27b, the Hilbert spectrum of the reconstructed signal not only extracts the rotating frequency 28 Hz (calculated value 25.02 Hz) but also indicates the characteristic frequency of inner race 125 Hz (calculated value 120.74 Hz), its second and third harmonic frequencies 246.9 Hz (calculated value 241.48 Hz) and 365.6 Hz (calculated value 362.22 Hz). Moreover, both of the characteristic frequency of inner race and its second harmonic frequency are surrounded by sidebands spaced at rotating frequency 25 Hz. Hence, the extracted features indicate that there existed a localized defect on the inner race of the front bearing. As comparison, the processing results derived by non-adaptive TQWT feature extraction method are displayed in Fig. 28, in which the fault features were not successfully identified.

6 Conclusions

Spectral kurtosis is an efficient and effective way to indicate impulsive transients masked by interfering components. The continuous development of wavelet decomposition theory significantly enhances the robustness of SK. In this chapter, we introduce the fundamentals of kurtogram, and focus on the recent advancements of fast kurtogram. QAWTF and SW are beneficial variations of discrete wavelet transform that exhibit non-dyadic time-frequency paving. With a range of engineering applications, the wavelet based kurtogram is verified to possess more robust noise-resistance and interference suppressing ability.

Acknowledgements The authors gratefully acknowledge the Project supported by the Natural Science Foundation of China (Grant No. 51605403), the Natural Science Foundation of Fujian Province, China(Grant No. 2016J01261), the Project supported by the Natural Science Foundation of Guangdong Province, China (Grant No. 2015A030310010)and for the financial and experimental support from Prof. Zhengjia He, Prof. Zhousuo Zhang, Prof. Yanyang Zi, all of whom are with school of mecanical engineering of Xi'an Jiaotong University, P.R. China.

References

1. Randall R. B. Vibration-based Condition Monitoring: Industrial, Automotive and Aerospace Applications, New Jersey: Wiely, 2011.
2. Wang Y. X., Xiang J. W., Markert R., Liang M., "Spectral kurtosis for fault detection, diagnosis and prognostics of rotating machines: A review with applications," Mechanical Systems and Signal Processing, 2016, 66–67:679–698.
3. J. Antoni, "The infogram: Entropic evidence of the signature of repetitive transients," Mechanical Systems and Signal Processing, 2015, 74:73–94.
4. Chen B. Q., Zhang Z. S., Zi Y. Y., et al. "Detecting of transient vibration signatures using an improved fast spatial–spectral ensemble kurtosis kurtogram and its applications to mechanical signature analysis of short duration data from rotating machinery," Mechanical Systems and Signal Processing, 2013, (40): 1–37.

5. Wang Y. X., Markert R., Xiang J. W., "Research on variational mode decomposition and its application in detecting rub-impact fault of the rotor system," Mechanical Systems and Signal Processing, 2015, 60–61:243–251.
6. Dwyer R. F., "Detection of non-Gaussian signals by frequency domain kurtosis estimation," in: International Conference on Acoustic, Speech, and Signal Processing, Boston, 1983, pp. 607–610.
7. Wang S. B., Chen X. F., Li G. Y., et al. "Matching demodulation transform with application to feature extraction of rotor rub-impact fault," IEEE Transactions on Instrumentation and Measurement, 2014, 63(5):1372–1383.
8. Otonnello C., Pagnan S., "Modified frequency domain kurtosis for signal processing," Electronics Letters, 1994, 30 (14):1117–1118.
9. Pagnan S., et al., "Filtering of randomly occurring signals by kurtosis in the frequency domain," Proceedings of the 12th International Conference on Pattern Recognition, vol. 3, October 1994, pp. 131–133.
10. Capdevielle V., Serviere C., Lacoume J. L., "Blind separation of wide-band sources: application to rotating machine signals," Proceedings of the Eighth European Signal Processing Conference, vol. 3, 1996, pp. 2085–2088.
11. Antoni J., Randall R. B., "The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines," Mechanical Systems and Signal Processing, 2006, 20: 308–331.
12. Antoni, J., "Fast computation of the kurtogram for the detection of transient faults," Mechanical Systems and Signal Processing, 2007, 21:108–124.
13. Antoni J., "The spectral kurtosis: a useful tool for characterising non-stationary signals," Mechanical Systems and Signal Processing, 20 (2006): 282–307.
14. Wang Y. X., Xiang J. W., Marckert R., Liang M., "Spectral kurtosis for fault detection, diagnosis and prognositcs of rotating machines: A review with applications," Mechanical Systems and Signal Processing,, 66–67 (2016) 679–698.
15. N. Sawalhi, R.B. Randall, "Spectral kurtosis optimization for rolling element bearings," in: Proceedings of the 8th International Symposium on Signal Processing and its Applications, 2005, 839–842.
16. Y.G. Lei, J. Lin, Z.J. He, Y.Y. Zi, "Application of an improved kurtogram method for fault diagnosis of rolling element bearings," Mechanical Systems and Signal Processing, 2011, 25 (5):1738–1749.
17. Wang. Y., Xiang J., Jiang Z., Lang L., "An adaptive SK technique and its application for fault detection of rolling element bearings," Mechanical Systems and Signal Processing, 2006, 20 (2):308–331.
18. J. Luo, Yu D., Liang M., "A kurtosis-guided adaptive demodulation technique for bearing fault detection based on tunable-Q wavelet transform," Measurement Science & Technology, 2013, 24 (5): 055009.
19. Zhang Z., Liang M.,Li C., Hou S., "Joint kurtosis-based adaptive band stop filtering and iterative autocorrelation approach to bearing fault detection," Journal of Vibration and Acoustics, 2013, 135(5): 051026.
20. Wang X., He Z., Zi Y., "Spectral kurtosis of multiwavelet for fault diagnosis of rolling bearing," Journal of Xi'an Jiaotong University, 2010, 44 (3):77–81.
21. Liu H., Huang W., Wang S., Zhu Z., "Adaptive spectral kurtosis filtering based on Morlet wavelet and its application for signal transients detection," Signal Processing, 2014, 96: 118–124.
22. Selesnick I. W., Baraniuk R. G., Kingsbury N., "The dual-tree complex wavelet transform - A coherent framework for multiscale signal and image processing," IEEE Signal Processing Magazine, 2005, 22 (6):123–151.
23. He W. P., Ding Y., Zi Y. Y., Selesnick. I. W., "Sparsity-based algorithm for detecting faults in rotating machines," Mechanical Systems and Signal Processing, 2016, 72:46–64.

24. Selesnick. I. W., “Wavelet transform with tunable Q-factor,” IEEE Transactions on Signal Processing, 2011, 59(8):3560–3575.
25. He W. P., Zi Y. Y., Chen B. Q., et al. “Tunable Q-factor wavelet transform denoising with neighboring coefficients and its application to rotating machinery fault diagnosis,” Science China Technological Sciences, 2013, 56 (8): 1956–1965.
26. He W. P., Zi Y. Y., Chen B. Q., et al. “Automatic fault feature extraction of mechanical anomaly on induction motor bearing using ensemble super-wavelet transform,”Mechanical Systems and Signal Processing, 2015, 54: 457–480.

Time-Frequency Manifold for Machinery Fault Diagnosis

Qingbo He and Xiaoxi Ding

Abstract In this chapter a new method called time-frequency manifold (TFM) is reported for signature enhancement and sparse representation of non-stationary signals for machinery fault diagnosis. In the framework of the TFM analysis, the phase space reconstruction is firstly employed to reconstruct the dynamic manifold embedded in an analysed signal, then the time-frequency distributions (TFDs) are generated in the reconstructed phase space to represent the non-stationary information, and manifold learning is finally addressed on the TFDs to discover intrinsic TFM structure. In this process, the TFM combines non-stationary information and nonlinear information simultaneously. This will provide a better time-frequency signature with the merits of noise suppression and resolution enhancement for machine health diagnosis. Furthermore, a TFM synthesis approach is further reported to explicitly recover the transient signal from the TFM signature by combining the sparse theory with the TFM structure. The objective of the introduced work is to exploit a TFM technology for enhancing the time-frequency signature and representing the transient feature with in-band noise suppression for machine fault signature analysis and transient feature extraction.

1 Introduction

Effective analysis of vibration signals is the basis of machinery fault diagnosis. However, in practice there always exists lots of background noise in collected vibration data, which will corrupt the fault-induced transient impulses. It is always an important aim to de-noise the measured vibration signal and extract the intrinsic fault signatures for a reliable fault diagnosis. During the past two decades, advanced signal processing techniques have been widely developed for effective machine fault feature extraction in machine health diagnosis area. Emerged techniques

Q. He (✉) · X. Ding
Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, Anhui, People's Republic of China
e-mail: qbhe@ustc.edu.cn

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors, Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_6

include wavelet transform [1], time-frequency distribution (TFD) [2], Hilbert-Huang transform (HHT) [3], multivariate statistical analysis [4] and spectral kurtosis [5], etc. Due to the nonlinear nature of machine vibrations, nonlinear features have attracted lots of attentions, e.g., fractal dimensions [6], complexity measure [7], phase space features [8], and kernel-based features [9], etc. In summary, nonlinear features can mainly be generated from two sources: one is the high-dimensional phase space data being reconstructed from raw signals, and the other is the information hidden in multi-dimensional vibration signals or multivariate features.

The signals generated by the defect-induced machine faults have the nature of nonstationarity for the varying frequency characteristics with time. For this type of signals, neither the time-domain techniques nor the frequency-domain ones can provide enough information for effective diagnosis due to their lack of information in each other domain. The time-frequency distribution (TFD) combines the information in time and in frequency, and is thus an effective approach for non-stationary signal analysis in machine fault diagnosis. Time-frequency representation can characterize varying frequency information at different time locations, providing plentiful non-stationary information of the analyzed signal. The TFD in the same health conditions will indicate a similar structure. If the machine condition changes, the TFD will also become significantly different. Hence, the TFD is beneficial to machine health diagnosis. However, the health pattern-related time-frequency structure will generally be corrupted by the other irrelevant components such as noise, which can worsen the resolution of useful time-frequency features. Many techniques have been addressed for de-noising the corrupted signals and extract the defective transient dynamics, such as, the time-domain averaging method [10], band-pass filtering [11], frequency-domain thresholding [12], empirical mode decomposition (EMD) [13], matching pursuit (MP) [14], orthogonal matching pursuit (OMP) [15]. These methods didn't focus on de-noising in the time-frequency domain, and thus may sacrifice part of the time-frequency resolution. Meanwhile, due to the lack of consideration for the relevancy and locality among the transient features of the signals, the in-band noise cannot be effectively removed from the raw transients while the pulse waveform structure is still well maintained.

Manifold learning has emerged in nonlinear feature extraction. It can identify low-dimensional nonlinear structure hidden in high-dimensional data, through several techniques including locally linear embedding (LLE) [16], isometric feature mapping (IsoMap) [17], and local tangent space alignment (LTSA) [18], etc. This merit has attracted increasing attentions in machine fault diagnosis area. Many studies have been focused on extracting nonlinear manifold features from high-dimensional system condition parameters, such as multi-dimensional time series in a reconstructed phase space [19], multivariate statistical feature extraction. In the sense of machine health diagnosis, manifold learning may be employed to reveal the change of machine health pattern through analyzing the inherent structure related to the nature of different faults. Different from traditional studies on manifold feature extraction, in this chapter, a TFM theory is thoroughly described by

well considering the merits of TFD and nonlinear features simultaneously [20, 21]. The TFM describes a kind of dynamic manifold of the time-frequency structure for the non-stationary signal. It can be extracted by a technique which addresses manifold learning on a series of TFDs in the reconstructed phase space [21]. The corresponding TFM signature represents an intrinsic time-frequency structure with a high resolution for representing impulse components of interest and excellent suppression effect for the noise. Due to its capability of combining the nonstationarity and the nonlinearity, this new signature is exactly suited for representing machine fault pattern and even demodulating the fault information in the time-frequency domain [22–24].

Furthermore, according to the merits of TFM analysis in noise suppression and resolution enhancement in the time-frequency domain [20, 21], it can be foreseen that it has the potential to solve the problem in TFA-based denoising approach. That is to say, theoretically the signals reconstructed from the TFM signature will have satisfactory denoising effects. Then a TFM synthesis approach was further reported for machinery fault signal denoising [25, 26]. The basic idea of this method is to synthesize a clearer TFD for the fault signal based on the TFM signatures of the raw signal by importing sparse theory into the TFMs. Based on the synthetic techniques, the clean fault signal can be synthesized again in the time domain. Hence, the recovered signal will have an excellent denoising performance in the application of machine transient feature extraction.

In this chapter, we present the theory of TFM technology for machine fault signature analysis and transient feature extraction, which aims to enhance the time-frequency signature and represent the transient feature with the in-band noise suppression. The TFM technology contains two aspects: TFM analysis and TFM synthesis. The TFM analysis mainly contributes on the time-frequency signature enhancement and noise suppression, but the waveform and amplitude of the fault signal will be seriously weakened. To overcome this shortcoming, the TFM synthesis is developed based on the TFM analysis and is used to extract and recover the transient feature. These two aspects will refine and extend the ability of the TFM technology in the application of machine health condition diagnosis.

2 Time-Frequency Manifold Analysis

2.1 Principle

It is known that the TFD, (which can be achieved by various TFA methods such as STFT, Continuous WT and Wigner-Ville distribution) can reveal the non-stationary pattern of a dynamic system. For samples under the same health condition, the TFD will display an intrinsic structure, which corresponds to a kind of nonlinear manifold. Nevertheless, the TFD will also indicate slight variance with the change of initial time and working condition, and can be seen as an output sample from the

underlying manifold. For a simulated bearing fault signal with −10 dB white noise, the TFDs of the noiseless signal and noisy signal are drawn in Fig. 1a, b respectively. It can be seen that the noise is distributed among the whole time-frequency plane including locations of the transient impulses. Meanwhile, it can be also found that basically the dynamic structure with the elliptical shape of TFD in Fig. 1c is submerged in the noise as displayed in Fig. 1d. The dynamic structure depicts the non-stationary information of the machinery health condition. For the noiseless signal, the dynamic structure can be seen as a manifold structure in the time-frequency domain, so here it is called the TFM. For the noisy non-stationary signal, the TFM is embedded on the TFD as an intrinsic nonlinear manifold structure in the time-frequency domain. For different vibration signals, the TFM corresponds to different time-frequency patterns. The TFM can be extracted by a technique which addresses manifold learning on a series of TFDs in the reconstructed phase space [20, 21]. This technique combines non-stationary information and nonlinear information together and can contribute a new TFD signature with the advantage of noise suppression. The basic idea of the TFM learning is to address manifold learning on reconstructed multi-dimensional TFDs of a non-stationary signal by implementing the following three steps: phase space reconstruction (PSR), TFD, and manifold learning. The following subsections describe the detailed theory of the TFM analysis technique.

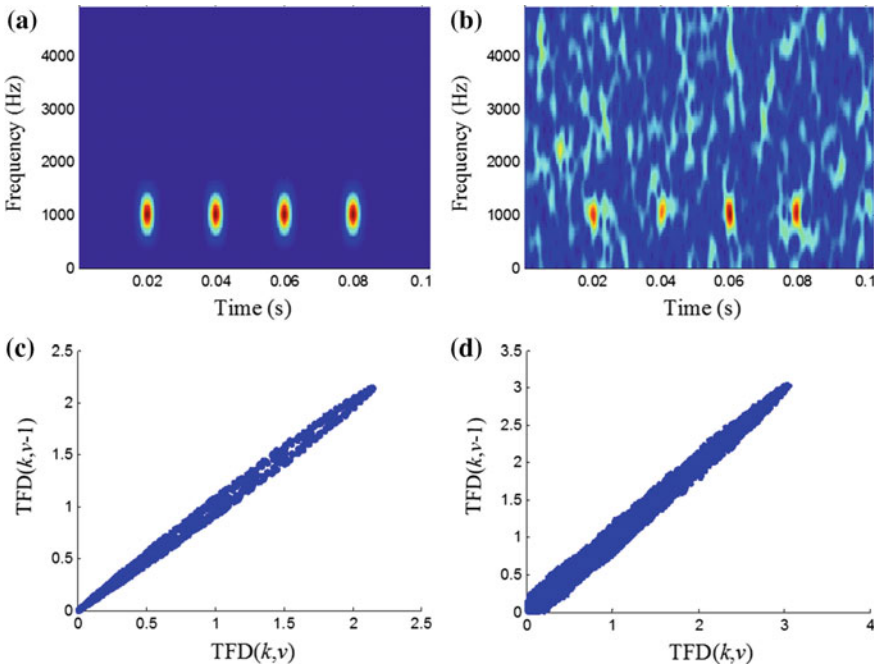


Fig. 1 Dynamic structure of the TFD: **a** the TFD and **c** the corresponding dynamic structure of the noiseless signal; **b** the TFD and **d** the corresponding dynamic structure of the noisy signal

2.2 Phase Space Reconstruction

The TFM learning requires firstly reconstructing the manifold of signal in a high-dimensional phase space by the phase space reconstruction (PSR) technique. The PSR is an effective method to reconstruct an inherent dynamic system that is embedded in an observed time series [27–29], which aims to trace out the orbit of the dynamic system in the reconstructed high-dimensional space. The manifold reveals the dynamic nature of the system, but is embedded in the observed time series. By applying this time-delay reconstruction, we can reconstruct the underlying manifold being embedded in a given signal. For a signal $x(t)$ with N data points, the ith phase point vector in an m-dimensional phase space is given as:

$$X_i^m = \left[x_i, x_{i+\tau}, \ldots, x_{i+(m-1)\tau}\right] \tag{1}$$

where x_i is the ith data point in $x(t)$, m is the embedding dimension, and τ is the time delay (an integer). For most real systems, the embedding dimension m is not a prior knowledge. There are many methods to calculate the dimension parameter, such as false neighbors [27] and the Cao's method [28]. In the process of determining the embedding dimension, the time delay could be set to be one by Takens' theory [29]. In this study, to keep a high time resolution for the TFM signature, the time delay should also be set to be one.

By the PSR, the phase point vectors $\{X_i^m \mid i = 1, 2, \ldots, n\}$ can be constructed after determining the embedding dimension m and the time delay τ. These vectors form an $m \times n$ matrix P ($\tau = 1$, $n = N - m + 1$) in phase space as below:

$$\begin{bmatrix} X_1^m \\ X_2^m \\ \cdots \\ X_i^m \\ \cdots \\ X_n^m \end{bmatrix}^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \\ x_2 & x_3 & \cdots & x_{m+1} \\ \cdots & \cdots & \cdots & \cdots \\ x_i & x_{i+1} & \cdots & x_{i+(m-1)} \\ \cdots & \cdots & \cdots & \cdots \\ x_n & x_{n+1} & \cdots & x_N \end{bmatrix}^T = \begin{bmatrix} P_x^1 \\ P_x^2 \\ \cdots \\ P_x^m \end{bmatrix} \tag{2}$$

According to Eq. (2), aligning the vectors $\{X_i^m \mid i = 1, 2, \ldots, n\}$ in the order of time produces m vectors $P_x^j \in R^{1\times n}, j = 1, 2, \ldots, m$, whose element indices correspond to the time. The time-series vectors $P_x^j (j = 1, 2, \ldots, m)$ are the rows of matrix P and can be considered as m signals denoted by $P_x^j(k) (j = 1, 2, \ldots, m)$ for convenience.

2.3 Time-Frequency Distribution

The TFD is an effective technique to analyze non-stationary signals [2]. It is employed to provide a 2-D representation that combines the information of time

and frequency for the constructed signal $P_x^j(k)$. In this paper, the STFT is taken to generate the TFD called the spectrogram. The spectrogram is a real-valued and non-negative energy distribution in the TF domain. Firstly, each row (with the time sense) of the data matrix P is analyzed by the STFT to provide a time-frequency representation as shown in the following equation:

$$S_j(k,v) = \sum_{l=-\infty}^{\infty} P_j[l]w[k-l]e^{-i\frac{2\pi}{M}vl}, \quad j = 1,2,\ldots,m \tag{3}$$

where k and v are the location of time axis and frequency axis, respectively, M is the discrete frequency points number in STFT, $w(k)$ is a short-time analysis window, and P_j is the jth row of matrix P with length n. The result $S_j(k, v)$ is in the complex form, which can be also expressed in two parts: amplitude A_j (k, v) and phase $\theta_j(k, v)$. The amplitude part is just the TFD of the analysed signal. Therefore, m TFDs can be generated from the constructed data P. As displayed in Fig. 1, the dynamic structure can be reconstructed in the phase space by PSR. Generally, the spectrogram-based TFD reveals a synthetic structure to evaluate machine faults. However, the background noise will also corrupt the TFD, which will hence worsen the resolution of time-frequency features related to the faults. This issue will be addressed by the following manifold learning technique.

In the following step, these m TFDs will be inputted into a manifold learning algorithm for TFM calculation to extract the dynamic structure. It should be noted that due to the nonlinear learning process of manifold learning for 3-D structure of m TFDs ($m \times$ 2-D TFD with the size of $L \times n$ in a high dimension, where L is the number of time points, n is the number of frequency points), there will be huge computation time cost in the process of TFM learning, which influences the use of the TFM in analyzing a long signal in diagnosis or multiple signals in monitoring. To improve the computational efficiency, a pre-processing technique needs to be employed to reduce the size of the calculated TFDs before TFM learning. Here, two simple but effective methods, including frequency band selection and two-dimensional discrete wavelet transform (2-D DWT), are suggested on the m TFDs to reduce the computational work. In the process of frequency band selection method, only the frequency band of interest is selected where the main vibration pattern can be revealed. For 2-D DWT, a low resolution "approximation" image can be obtained for manifold learning based on the multi-resolution analysis theory. Through the pre-processing techniques, m TFDs (the size of each TFD is $L' \times n'$, where L' is less than or equal to L and n' is less than n) are generated by the constructed data $P_x^j(k)(j = 1,2,\ldots,m)$ in the phase space.

2.4 TFM Learning

With the PSR and TFD, all of m TFDs form a 3-D matrix with the size of $m \times L' \times n'$. As demonstrated in Fig. 2a, the original TFD of the simulated signal suffers a severe problem in time-frequency resolution due to noise corruption. This problem has also been brought into the phase space as shown in Fig. 2b, which displays the 3-D structure of m TFDs. To solve the problem above, the manifold learning is employed. The basic idea is illustrated in Fig. 2c. Due to the time-delay operation, each TFD in the phase space also results in a time-delayed representation. But note that the time delay of the TFDs is not rigorous. This is because the Fourier transform operation in each short-time frame would cover different time information for different constructed signals in phase space. The TFM intends to reveal varying time-frequency characteristics among these multi-dimensional signals as illustrated in Fig. 2c. The revealed solid time-frequency structure corresponds to a specific function contained in a specific range of time, and can be imagined as a skeleton form, e.g., as shown in Fig. 1c. By the manifold learning in the phase space, the deterministic information (e.g., the fault information) will be kept for its inherent manifold form; however, the random information (e.g., the noise) will not be equally processed as there is not a solid skeleton structure for the noise (although they are partly correlated) among different TFDs. This is just the motivation to study the TFM from the time-delayed TFDs for machine fault signature representation.

Manifold learning aims at discovering the intrinsic and dynamic structure of the TFM from the generated m TFDs. Since these TFDs form a 3-D matrix with the size

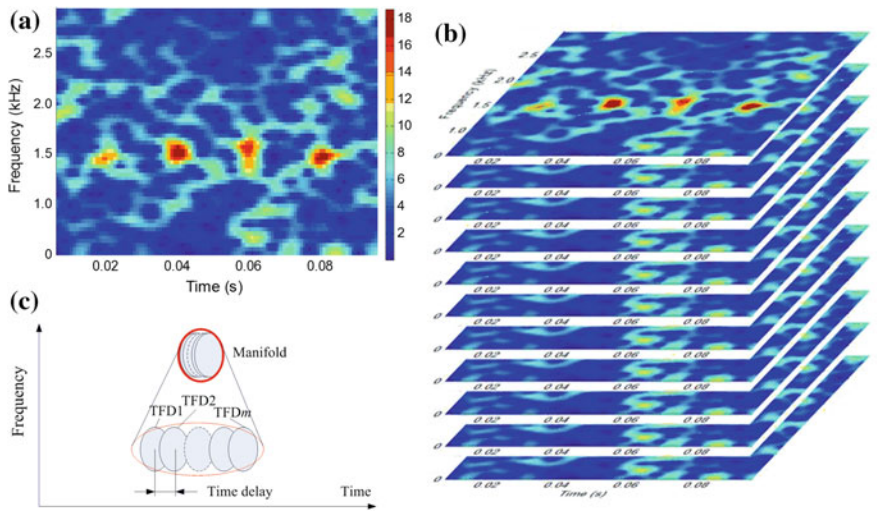


Fig. 2 Demonstration of TFM learning: **a** original TFD, **b** m TFDs in reconstructed phase space, and **c** illustration of TFM learning theory

of $m \times L' \times n'$, we should change the matrix to a 2-D one to satisfy the input of manifold learning algorithm. This is realized by organizing each 2-D TFD matrix to a 1-D vector by connecting one column by one column. The manifold learning is then performed on the reorganized TFDs matrix with the size of $m \times (L' \times n')$ This study employs the LTSA algorithm, which is to construct an approximation for the tangent space at each data point, and then align those tangent spaces to obtain the global coordinates of the data points with respect to the underlying manifold. Details and derivation of the algorithm can be found in [18]. The following gives a simple description of the algorithm for the TFM learning.

Given a data set $Z = [z_1, \ldots, z_{(L' \times n')}] \in R^{m \times (L' \times n')}$ with $z_i \in R^m$, sampled (possibly with noise) from a d-dimensional TFM $d \ll m$. The data set Z represents the TFD pixels in the m-dimensional phase space. The LTSA algorithm produces $(L' \times n')$ d-dimensional coordinates $T \in R^{d \times (L' \times n')}$ for the manifold constructed from a series of local nearest neighbors. The algorithm is conducted in the following steps.

Step 1 Local neighborhood construction

For each $z_i, i = 1, \ldots, (L' \times n')$, determine its k nearest neighbors $z_{i_j}, j = 1, \ldots, k$, and form a neighborhood set $Z = [z_1, \ldots, z_{(L' \times n')}]$.

Step 2 Local linear fitting

Compute the orthonormal basis matrix Q_i for the d-dimensional tangent space at z_i based on the local neighborhood Z_i. It can be taken as the matrix of d left singular vectors of $Z_i(\mathrm{I} - \mathrm{ee}^T/k)$ corresponding to its d largest singular values through the singular value decomposition (SVD) as $Z_i(\mathrm{I} - \mathrm{ee}^T/k) = Q_d \sum_d V_d^T$, where e is a vector of all ones. Each data point z_{i_j} in the neighborhood of z_i is then projected to the computed tangent space as $\theta_j^{(i)} = Q_i^T(z_{i_j} - \bar{z}_i)$, where $\bar{z}_i$ is the mean of z_{i_j}'s. Then we can get $(L \times n')$ local coordinates $\Theta_i = [\theta_1^{(i)}, \ldots, \theta_k^{(i)}]$, $i = 1, \ldots, (L \times n')$.

Step 3 Local coordinates alignment

Align the $(L \times n')$ local projections$\Theta_i = [\theta_1^{(i)}, \ldots, \theta_k^{(i)}]$, $i = 1, \ldots, (L' \times n')$, to obtain the global coordinates g_i, $i = 1, \ldots, (L' \times n')$. Denote $T_i = [g_{i_1}, \ldots, g_{i_k}]$ with the index set $\{i_1, \ldots, i_k\}$ determined by the neighbors of each z_i. Let $E_i = T_i(\mathrm{I} - \mathrm{ee}^T/k) - L_i\Theta_i$ be the local reconstruction error matrix, where L_i is the local affine transformation matrix. To minimize the local reconstruction error, the optimal alignment matrix L_i is given by $L_i = T_i(\mathrm{I} - \mathrm{ee}^T/k)\ \Theta_i^+ = T_i\Theta_i^+$, where Θ_i^+ is the Moore-Penrose generalized inverse of Θ_i. Then E_i can be written as

$$E_i = TS_iW_i \tag{4}$$

where S_i is the $0 - 1$ selection matrix such that $TS_i = T_i$ and $W_i = (\mathrm{I} - \mathrm{ee}^T/k)(\mathrm{I} - \Theta_i^+\Theta_i)$. The single data set manifold alignment of LTSA is achieved by minimizing the following global reconstruction error:

$$\sum_i \|E_i\|^2 \equiv \sum_i \|TS_iW_i\|_F^2 = \|TSW\|_F^2 \tag{5}$$

where $S = [S_1, \ldots, S_{(L\times n')}]$ and $W = \mathrm{diag}(W_1, \ldots, W_{(L'\times n')})$. To uniquely determine T, impose $TT^T = \mathrm{I}_d$. The alignment matrix can be formed as $B = SWW^TS^T$ to solve the optimal problem.

Step 4 Aligning global coordinates

Compute the $d + 1$ smallest eigenvectors of B, and pick up the eigenvector matrix $[u_2, \ldots, u_{d+1}]$ corresponding to the $2nd$ to the $d + 1$ smallest eigenvalues, and set $T = [u_2, \ldots, u_{d+1}]^T$. The results of $T(\in R^{d\times(L'\times n')})$ then correspond to the global coordinates of the low-dimensional TFMs. The TFM matrix can be reorganized to be a 3-D matrix with the size of $d \times L' \times n'$. The 3-D TFM structure can be denoted by $M_{tf}^d(k, v)$. Each dimensional TFM signature has the same appearance as the TFD. Note the dimension d is far less than the phase space dimension m. Meanwhile, as demonstrated in Fig. 3a, the learned TFM signature for the simulated signal extracts the time-frequency structure with noise suppression and resolution enhancement. And the corresponding dynamic structure embedded in the TFM signature as displayed in Fig. 3b is obviously improved with elliptical shape, which is similar to the dynamic structure in Fig. 1c.

2.5 Procedure of Time-Frequency Manifold Analysis

In summary, the procedure of the proposed TFM analysis can be described briefly as follows:

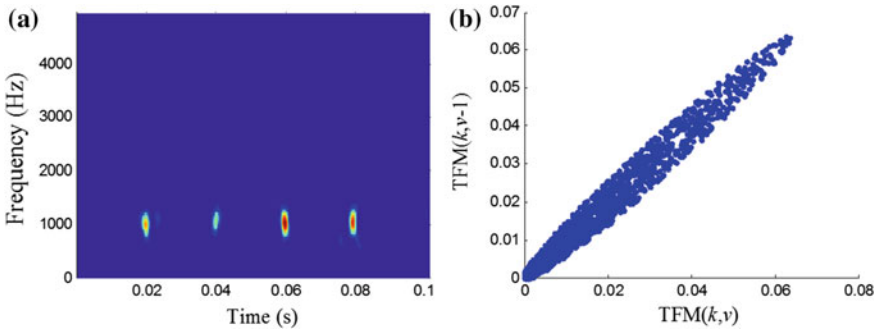


Fig. 3 **a** The TFM and **b** the corresponding dynamic structure

Step 1 Given a signal $x(t)$ with N data points, calculate the data matrix P of size $m \times n$ ($n = N - m + 1$) by PSR according to Eq. (2).
Step 2 Do the STFT for each row of matrix P via Eq. (3) to get $S_j(k, v)$, $j = 1,2, \ldots, m$, and calculate the corresponding amplitude A_j (k, v) and phase $\theta_j(k, v)$.
Step 3 Do the pre-processing technique to get m TFDs with the size of $L' \times n'$.
Step 4 Calculate the TFM signature $M_{tf}^{d}(k, v)$ of size $L' \times n'$ by the LTSA algorithm.

The TFM structures $M_{tf}^{d}(k, v)$ reflect the time-frequency nature of system health condition by combining the nonstationarity and the nonlinearity. So the new structures are exactly suited for representing the system health signature for machinery fault diagnosis, which will be verified in the following Sect. 4.

3 Time-Frequency Manifold Synthesis

Motivated by the benefit of TFM in noise suppression, the elements of T can be applied as the principal sparse components, which are unit orthogonal vectors orderly corresponding to the d smallest eigenvalues in the LTSA algorithm. And the manifold T represents the nonlinear structure pattern of the original signals. Thus, this section reports an approach called TFM synthesis for signal denoising and transient signal detection, which mainly combines the techniques of TFD re-generation, time-frequency synthesis and PSR synthesis. The new principle of data denoising aims to reduce background noise of signals effectively, and at the same time keep the essence of transient signals to the maximum extent. Therefore, the proposed method is especially suited for denoising of machinery faulty vibration signals.

3.1 Principle

As manifold learning can keep the intrinsic nonlinear structure in dimension reduction of a high-dimensional data matrix, the TFM signature represents the time-frequency structure nature of the original signal in the sense of noise suppression, which can exhibit excellent denoising performance in the time-frequency domain. However due to the nonlinear process of the LTSA algorithm (the principle in Sect. 2.4), the TFM signature that is organized from the global coordinates T learned by LTSA will loss the amplitude information as compared to the TFD of the original signal. Mathematically, the LTSA is an optimal alignment algorithm by solving an eigenvalue problem. In order to uniquely determine the global coordinates T (the physical meaning is to reveal the intrinsic time-frequency structure), the following constraint condition is imposed:

$$TT^T = \mathrm{I}_d \tag{6}$$

where each row of T corresponds to a TFM signature and d is the given dimension. The eigenvector $T_i (i \in [1, d])$ is thus the TFM with a similar structure to the input TFD. Therefore, one of the most important issues is how to recover the amplitude information in the time domain.

3.2 TFD Re-Generation

In the principle of TFM synthesis, the m original TFDs are represented by the achieved TFM patterns respectively. That is to say, the principal time-frequency structure of the transient impulses in the raw signal can be learned by TFM. Hence, the m original TFDs can be realized again by a few learned TFMs of T as follows:

$$\hat{A}_j = \sum_{l=1}^{K} c_{j,l} M_{tf}^{l} \quad (K \leq d, j = 1, 2, \ldots, m) \tag{7}$$

where M_{tf}^{l} is the lst TFM matrix corresponding to the vector component T_l and $c_{j,l}$ is the corresponding sparse coefficient. According to Eq. (6), the sparse coefficient $c_{j,l}$ can be calculated from an inner product operation (sum of dot product of two matrices) between the original jth TFD matrix A_j and the lth TFM matrix M_{tf}^{l}.

Since the dynamic information and transient feature are mainly extracted and kept in the first or two TFMs, for simplification, only the first manifold $M_{tf}^{1}(k, v)$ represented as T_1 is employed in this process of TFM synthesis in this chapter. Thus, for each TFD of the vector P_x^j (j = 1, 2, …, m), the synthesis can be rewritten as follows:

$$\hat{A}_j = c_{j,1} M_{tf}^{1} \quad (j = 1, 2, \ldots, m) \tag{8}$$

By treating the TFM signature as a processed time-frequency base, the original amplitude A_j (k, v) can be re-generated as $\hat{A}_j$ (k, v), which can keep the intrinsic time-frequency structure while the random noise can be restrained. Moreover, it should be noted that, corresponding to the pre-processing techniques as mentioned in the TFM learning of Sect. 2.4, the learned TFM signature with the size of $L' \times n'$ will finally receive a matrix zero padding for the frequency band selection case, or a 2-D inverse DWT (2-D IDWT) to recover the same size of $L \times n$ with the original TFD.

3.3 Time-Frequency Synthesis

Furthermore, as the original phase $\theta_j(k, v)$ keeps the information of waveform structures in the raw signal, combining the re-generated amplitude $\hat{A}_j$ (k, v) with the

original phase information will improve the description errors of the re-generated signal compared to the real vibration signal. Then, m updated STFT results, denoted by $\hat{S}_j(k, v)$, $j = 1,2, \ldots, m$, can be generated. Then time-frequency synthesis is employed on each updated STFT result to calculate a new data matrix $\hat{P}$ in the phase space. The time-frequency synthesis of STFT can be expressed as follow:

$$\hat{P}_j[k] = \frac{1}{Mw[0]} \sum_{v=0}^{M-1} \hat{S}_j(k, v) e^{i\frac{2\pi}{n}kv}, \; j = 1, 2 \cdots, m \tag{9}$$

Assume $w[n]$ (with the window band width w_c and the window length N_w) is not equal to zero in the limited window length, then Eq. (9) holds under the following constraints [14]: (a) the time sampling factor L of STFT should satisfy the Nyquist criteria, that is $L \leq \frac{2\pi}{wc}$; (b) the frequency sampling interval should satisfy $\frac{2\pi}{N} \leq \frac{2\pi}{Nw}$, that is $N \geq N_w$. As each STFT result can generate a time series by time-frequency synthesis, m time series could thus construct a data matrix $\hat{P}$ with the same size as original data matrix P.

3.4 PSR Synthesis

After getting the updated data matrix $\hat{P}$ in the phase space, the PSR synthesis is applied to reconstruct the signal with denoising effect. In the process of reconstruction, it should be considered that every element of the original time series may appear at several places in the phase space data matrix. The PSR synthesis is presented as follows to reconstruct the signal from data matrix $\hat{P}$:

$$\hat{x}_i = \frac{\sum_{q \in \{I_i(j,k)\}} \hat{P}_q}{C_i}, \; i = 1, 2 \cdots, N; j = 1, 2 \cdots, m \tag{10}$$

where $\{I_i(j, k)\}$ is the subscript set of the signal elements that meets the requirement of $k + (j - 1)\tau = i$ ($k \in [1, N - (m - 1)\tau]$), and C_i is the number of elements in $\{I_i(j, k)\}$. The final result of the denoised signal can be thus represented as $\hat{x}(t)$ with N data points.

According to the principle of TFM synthesis, the denoised results of the simulated signal not only inherit the merits of TFM but also further improve the waveform shape of the transients as drawn in Fig. 4a. Moreover, as compared to the dynamic structures in Figs. 1c and 3b, the amplitude of the corresponding dynamic structure can be recovered as shown in Fig. 4b. Thus, the TFM synthesis is an advanced TFM analysis which is effective for signal denoising and transient feature extraction.

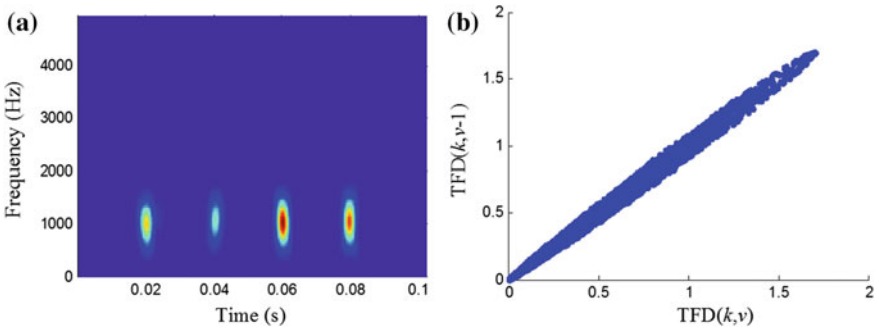


Fig. 4 The denoised signal based on TFM synthesis: **a** the TFD and **b** the corresponding dynamic structure

3.5 Procedure of TFM Synthesis

In summary, the procedure of the proposed TFM Synthesis can be described briefly as follows:

Step 1 Obtain the TFM signature according to the procedure of TFM learning in Sect. 2.5.

Step 2 Re-generate the amplitude A_j (k, v) $(j = 1,2, \ldots, m)$ of the original TFD as $\hat{A}_j$ (k, v) based on Eq. (8).

Step 3 Update the STFT results using the original phase part $\theta_j(k, v)$ and reconstructed amplitude $\hat{A}_j$ (k, v) to get $\hat{S}_j(k, v)$, $j = 1,2, \ldots, m$.

Step 4 A new data matrix $\hat{P}$of size $L \times n$ in phase space is generated by time-frequency synthesis according to the Eq. (9).

Step 5 The denoised signal $\hat{x}(t)$ is finally reconstructed by PSR synthesis by Eq. (10).

The reconstructed signal inherits the merits of TFM in well representing the non-stationary information with good in-band noise suppression. Thus the new manifold synthesis approach is quite suitable for machine transient feature extraction, which will be verified in the next Sect. 4.

Moreover, based on the merits of TFM analysis in noise suppression and resolution enhancement in the time-frequency domain, the sparse representation can be also employed on the TFM signature to learn the sparse components for transient feature extraction. In this way, the original TFD can be synthesized in the view of sparse expression. This will bring much better denoising effect for machinery fault diagnosis [30].

4 Experiments for Machinery Fault Diagnosis

The defect-induced rotating machine fault contains three typical characteristics: periodic due to the rotating nature, impulsive as induced by the defect, and transient in the amplitude. These defect-induced faults include the breakage of gear teeth, defects of typical bearing components, etc. However, the fault information is usually buried in heavy noise in measured vibration signals. With the capability of combining the nonstationarity and the nonlinearity, the TFM analysis is expected to effectively reveal the underlying time-frequency structure related to the defect-induced fault. Moreover, the TFM synthesis is used to recover the transient features, as well as inherent time-frequency structure keeping. To verify the effectiveness of the proposed TFM analysis and TFM synthesis for machinery fault diagnosis, the applications to gear fault diagnosis and bearing defect diagnosis are studied in the following.

4.1 Gear Fault Diagnosis

The experimental data was generated from an automobile transmission gearbox, which has 5 forward speeds and one backward speed as shown in Fig. 5. Vibration signals were acquired by using an accelerometer mounted on the outer case of the gearbox when it was loaded on the third speed. During a fatigue test, a tooth-broken fault occurred at the driving gear of the third speed at the beginning of Cycle 7. Thus Cycle 6 corresponds to the severe wearing fault stage. In the experiment, the input rotating speed was 1600 rpm and the sampling frequency was set at 3000 Hz. The working parameters of the third speed are shown in Table 1, where the

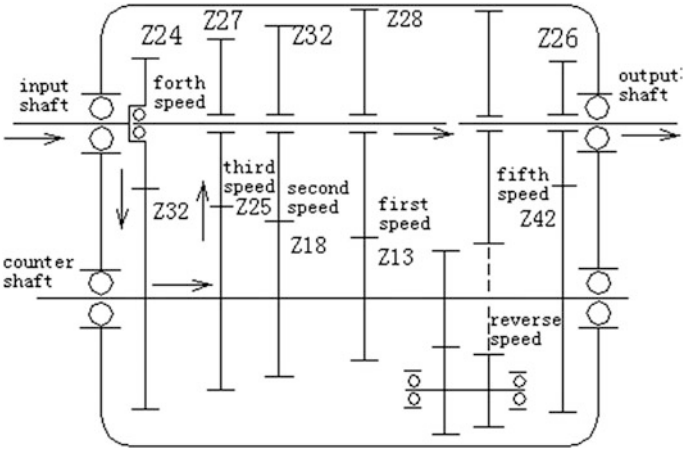


Fig. 5 Structure of the automobile transmission gearbox

Table 1 Working parameters of the third speed gears

Gear	Number of teeth	Rotating frequency (Hz)	Meshing frequency (Hz)
Driving gear	25	20	500
Driven gear	27	18.5	500

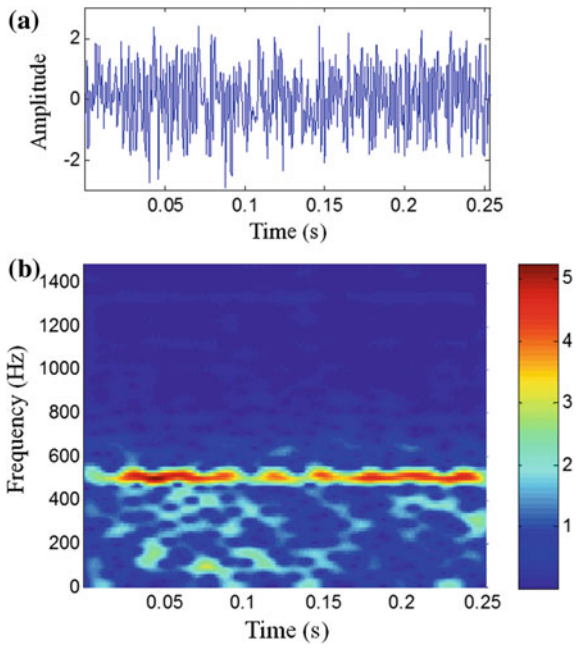


Fig. 6 The healthy gear signal: **a** waveform and **b** spectrogram

meshing frequency and the rotating frequency of the tested gear are calculated to be 500 Hz and 20 Hz, respectively.

In this section, the healthy and severe wearing faulty signals are taken for further analysis. The healthy signal as shown in Fig. 6 is first analyzed. The original spectrogram shows a synthetic but also corrupted result with heavy noise and couldn't tell when the frequency component occurs. However, as seen from Fig. 7, the first TFM signature shows a much clearer result in comparison with the original TFD. It can be found that the TFM signature for the healthy signal just contains the frequency component of 500 Hz that exists at any time, corresponding to the meshing frequency of the tested gear. In addition, based on TFM synthesis, Fig. 8a describes an irregular volatility for the healthy wearing signal. Figure 8b gives a much clearer TFD with only one meshing frequency band which is similar to the ones in Fig. 6b. Therefore, this result confirms that the proposed denoising method can depict the health condition of the tested gear.

For the severe wearing faulty signal, the fault signature cannot be clearly identified in the time domain as shown in Fig. 9. The spectrogram as shown in Fig. 9b represents a combination of time and frequency information, which can tell

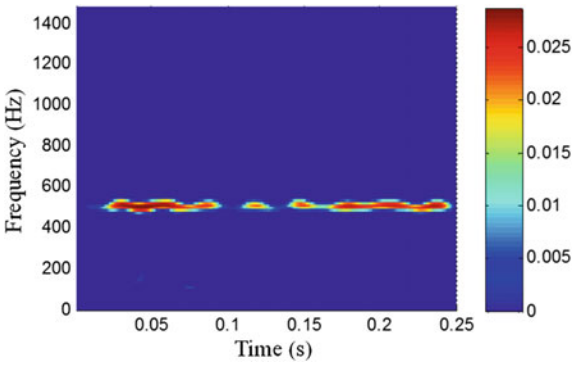


Fig. 7 The first TFM signature of the healthy gear signal

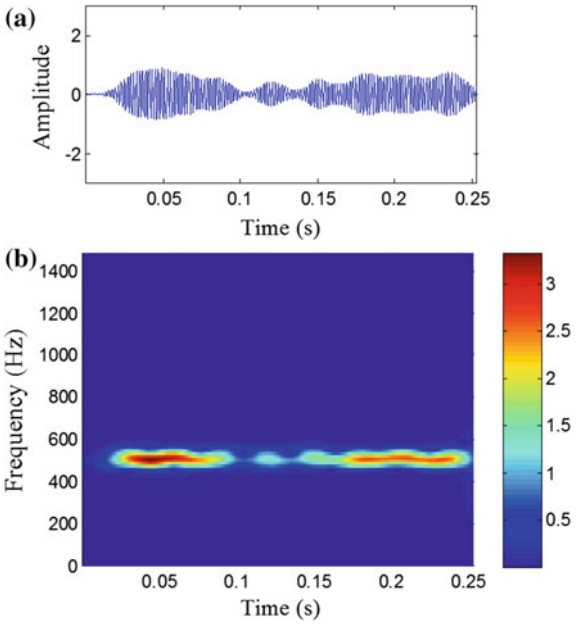


Fig. 8 The denoised healthy gear signal based on TFM synthesis: **a** waveform and **b** spectrogram

how the frequencies happen with the time. However, it can be seen that the time-frequency signature is corrupted by background noise. Then the proposed TFM analysis is employed to extract the time-frequency nature reflecting system health condition. As plotted in Fig. 10, it can be seen that the extracted TFM signature shows a rather clearer representation on the non-stationary structure in comparison with the original spectrogram. Meanwhile, the first TFM signature shows two typical frequencies, 500 and 280 Hz. The frequency of 500 Hz exists at any time locations corresponding to the meshing frequency but has a relatively weak energy. The frequency of 280 Hz appears periodically in a certain interval of

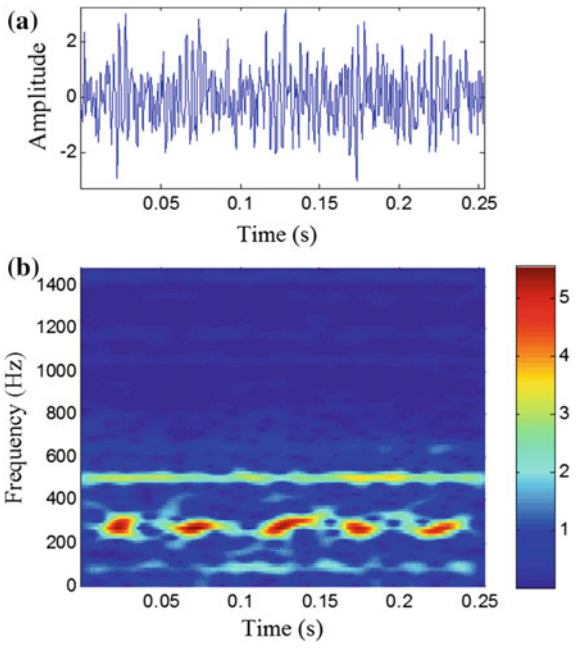


Fig. 9 The severe wearing faulty gear signal: **a** waveform and **b** spectrogram

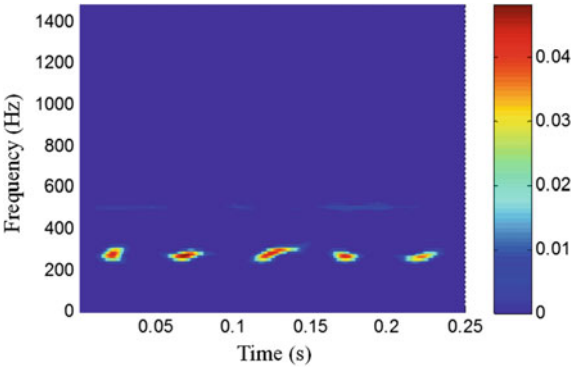


Fig. 10 The first TFM signature of the severe wearing faulty gear signal

about 0.05 s, which reflects the appearance of periodic impulses. Therefore, the first TFM signature can be judged as the existence of the severe wearing fault on the tested gear.

To further investigate the effect of the transient feature extraction, the TFM synthesis is applied to re-generate the original TFD based on the TFM signature. The reconstructed signal is shown in Fig. 11. It can be seen that the waveform of the denoised signal has much less noise and the periodic impulses could be identified effectively, which shows a good sparse property of periodic impulses. Meanwhile, the natural TF structure of the impulses is very well retained with the

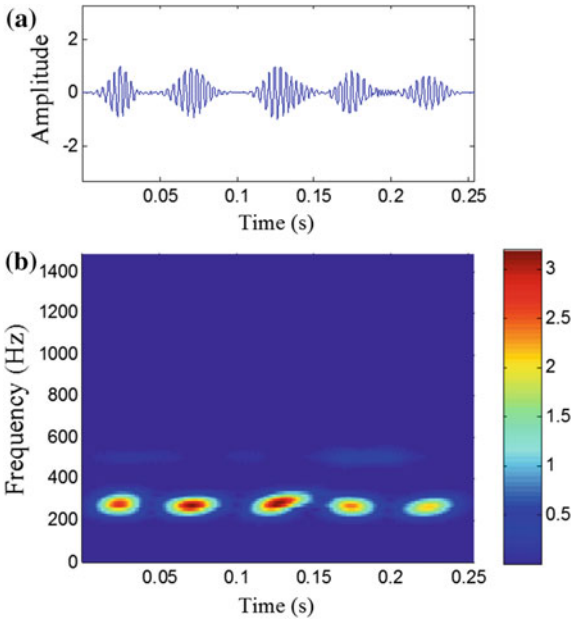


Fig. 11 The denoised severe wearing faulty gear signal based on TFM synthesis: **a** waveform and **b** spectrogram

in-band noise being greatly reduced in the TFD of the reconstructed signal. The result confirms that the TFM synthesis can effectively maintain the original periodic impulse structure and the meshing components while reducing noise, which is suitable for gear fault diagnosis.

4.2 Bearing Defect Diagnosis

The experimental data were collected by an experimental setup as shown in Fig. 12 from Case Western Reserve University Bearing Data Center [31]. Single point faults were set on the testing drive-end bearings (deep groove ball bearing with the Type 6205-2RS JEM SKF) separately at the rolling element and outer raceway using electric discharge machining method. The resulting vibration from the motor was measured by accelerometers being mounted to the motor's shell with magnetic bases, at a sampling frequency of 12 kHz. In this case study, the single-fault bearing with outer-race and rolling-element defect were analyzed. The outer-element defective bearing was at the rotating speed 1749 rev/min with 0.11×0.014 inches depth while the rolling-element defective bearing was at the rotating speed 1796 rev/min with 0.11×0.021 inches depth. The defective characteristic frequencies can be calculated to be 104.5 Hz and 141.1 Hz, respectively.

In the following, the defective signals with outer-race and rolling-element defect types are analyzed by the TFM analysis and TFM synthesis, respectively. In this chapter, the parameters for PSR are set to be $m = 13$ and $\tau = 1$ in the process of

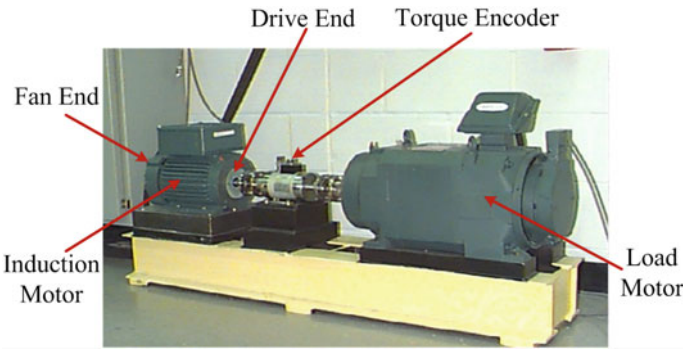


Fig. 12 The bearing test stand

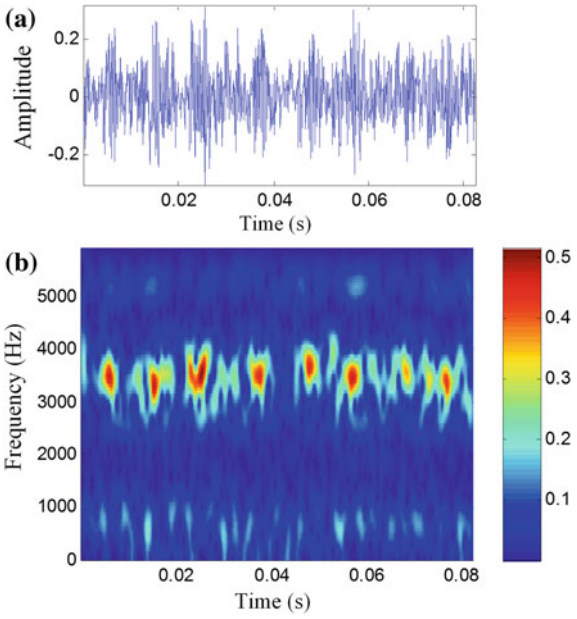


Fig. 13 The outer-race defective bearing signal: **a** waveform and **b** spectrogram

TFM learning. Figure 13 shows the waveform and the TFD of the raw outer-race defective bearing signal. It can be seen that in the waveform there are a series of periodic impulses submerged in the noise. Although the TFD presents a combination of time and frequency information, the noise corruption exists in a wide frequency band. This will influence the diagnostic performance. As drawn in Fig. 14, the achieved TFM signature shows an excellent result for the defect-induced fault signature with a good time-frequency resolution. The TFM reveals that the defect was located on the outer raceway of the tested bearing. In this process, the TFM analysis shows its effectiveness in extracting the underlying structure of the defect-induced fault.

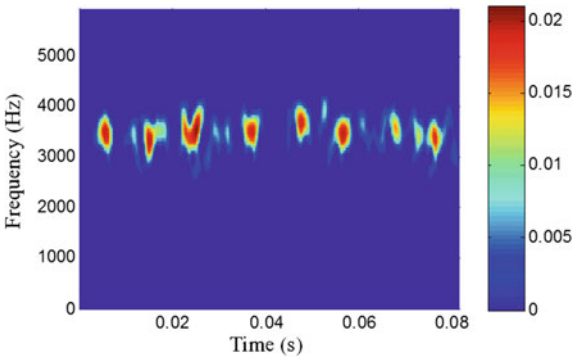


Fig. 14 The first TFM signature of the outer-race defective bearing signal

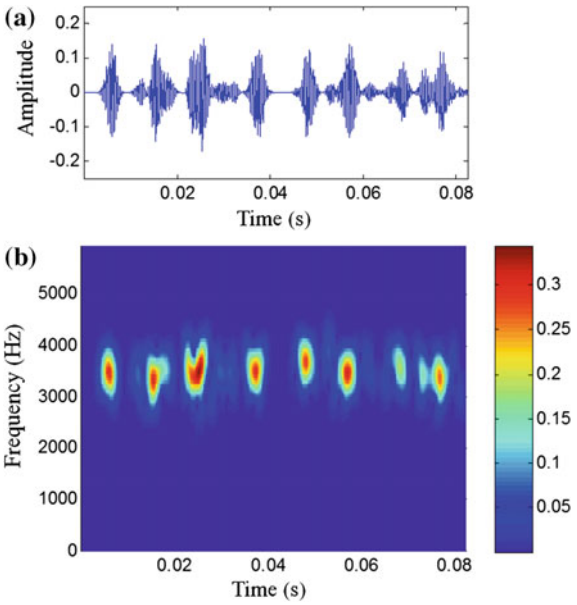


Fig. 15 The denoised outer-race defective bearing signal based on TFM synthesis: **a** waveform and **b** spectrogram

However, it can be also easily found that the extracted impulses lost its amplitudes compared to the raw ones in Fig. 13b. Due to the assumption $TT^T = \mathrm{I}_d$ in the nonlinear manifold learning process as introduced in Sect. 2.4, the TFM result will lose the available amplitudes heavily. Moreover, the shapes of the raw time-frequency transients are pruned to some extent. Therefore, the TFM synthesis is used to recover the transients of the raw signal with the amplitudes recovered and waveforms remained. From Fig. 15a, it can be seen that the impulses are much clearer than those in Fig. 13a with a high SNR and the in-band noise has been very well removed. At the same time, comparing the TFDs in Figs. 13b, 14 and 15b, the TFD of the synthetic signal clearly shows the merits in amplitude recover and

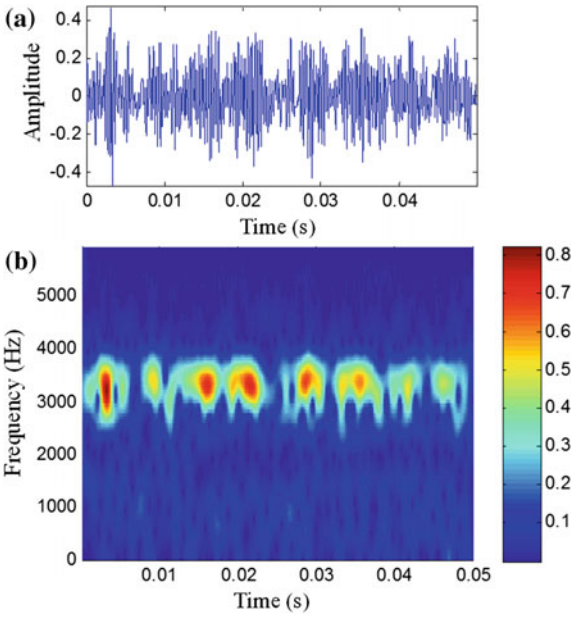


Fig. 16 The rolling-element defective bearing signal: **a** waveform and **b** spectrogram

waveform remaining. Therefore, the proposed denoising method is verified to be able to reduce noise effectively, as well as to keep the nature of fault signals.

Next, the rolling-element defective signal is analyzed by the proposed method. The waveform and the TFD of the defective signal are shown in Fig. 16. It can be seen that the waveform indicates a series of similar periodic impulses along the time, but the repetitive period of impulses cannot be identified from the waveform as there is noise corruption. Although the original spectrogram shows an obvious concentration of energy at the band near 3200 Hz, the noise being contained in this band reduces the resolution of the defect-induced impulses. The TFM signature as shown in Fig. 17 only keeps the natural structure related to the fault, and thus shows a rather clearer result for emphasizing the fault characteristics. Therefore, the TFM signature reveals the exact defect physics for the rolling element defect with its merits in enhancing the time-frequency resolution. The waveform and its TFD of the denoised signal based on TFM synthesis are demonstrated in Fig. 18. It can be easily seen that the transient features are much more obvious with regularity in Fig. 18a and the time-frequency impulses remain the shapes as the original ones in Fig. 16b in a certain degree, although the synthesis only recovers part of the amplitudes. This performance will be further improved by employing more TFMs to synthetize the original m TFDs. Thus, the result confirms that the proposed denoising method can reduce noise effectively and keep the natural structure of fault signals.

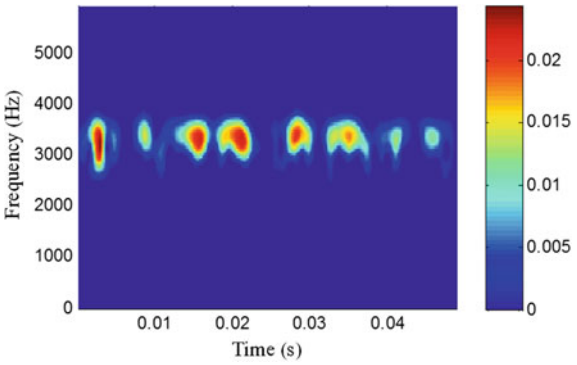


Fig. 17 The first TFM signature of the rolling-element defective bearing signal

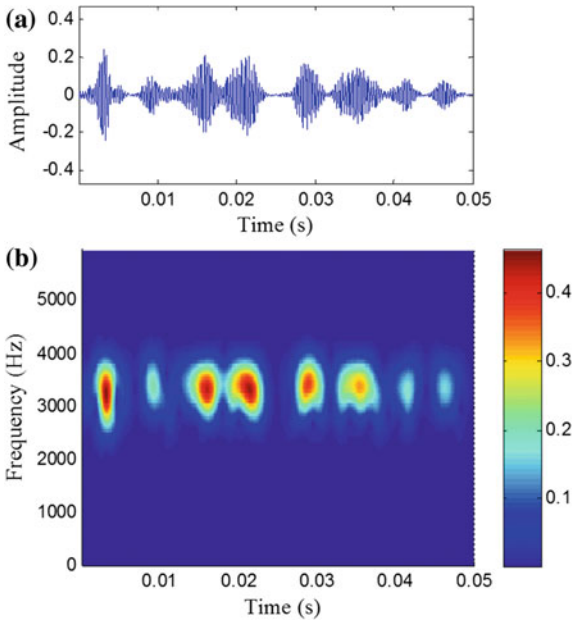


Fig. 18 The denoised rolling-element defective bearing signal based on TFM synthesis: **a** waveform and **b** spectrogram

5 Conclusions

This chapter describes a new TFM analysis approach by combining the time-frequency analysis and the nonlinear manifold for a better representation of machine health pattern. The TFM signature can reveal an intrinsic time-frequency structure related to the machine health pattern, and hence it is especially suited for analyzing the non-stationary fault signature of rotating machines. Motivated by the merits of the TFMs in high-resolution time-frequency signature analysis, a TFM synthesis approach is further presented for the time-domain information representation in the application of transient feature extraction. The TFM synthesis inherits

the merits of TFM signature in noise suppression and resolution enhancement to represent an intrinsic time-frequency structure, so it does not only reduce background noise effectively, but also keeps the intrinsic time-frequency structure of the periodic transient impulses. This is significant for intrinsic vibration data characteristics and reliable fault diagnosis. In conclusions, the TFM analysis provides an effective time-frequency approach to contribute useful features for fault diagnosis and the TFM synthesis builds a helpful approach to extract transient feature for signal denoising. Experimental studies on machine fault signature analysis (including gear fault diagnosis and bearing defect diagnosis) show the excellent merits of the TFM analysis and TFM synthesis, which indicates a potential of the TFM theory in enhancing time-frequency signature analysis for practical engineering applications.

References

1. Peng, Z. K. and Chu, F. L., "Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography", Mechanical Systems and Signal Processing, 2004, 18(2): 199–221.
2. Hammond J K and White P R., "The analysis of non-stationary signals using time-frequency methods", Journal of Sound and Vibration, 1996, 190(3): 419–447.
3. Yan R. and Gao R. X., "Hilbert–Huang transform-based vibration signal analysis for machine health monitoring", IEEE Transactions on Instrumentation and measurement, 2006, 55(6): 2320–2329.
4. Malhi A. and Gao R X., "PCA-based feature selection scheme for machine defect classification", IEEE Transactions on Instrumentation and Measurement, 2004, 53(6): 1517–1525.
5. Wang Y., Xiang J., Markert R., and Liang M., "Spectral kurtosis for fault detection, diagnosis and prognostics of rotating machines: A review with applications", Mechanical Systems and Signal Processing, 2016, 66: 679–698.
6. Feng Z, Zuo M J. and Chu F., "Application of regularization dimension to gear damage assessment", Mechanical Systems and Signal Processing, 2010, 24(4): 1081–1098.
7. Yan R. and Gao R X., "Complexity as a measure for machine health evaluation", IEEE Transactions on Instrumentation and Measurement, 2004, 53(4): 1327–1334.
8. Wang G F, Li Y B. and Luo Z G., "Fault classification of rolling bearing based on reconstructed phase space and Gaussian mixture model", Journal of Sound and Vibration, 2009, 323(3): 1077–1089.
9. He Q., Kong F. and Yan R., "Subspace-based gearbox condition monitoring by kernel principal component analysis", Mechanical Systems and Signal Processing, 2007, 21(4): 1755–1772.
10. Braun S., "The synchronous (time domain) average revisited, Mechanical Systems and Signal Processing", 2011, 25(4): 1087–1102.
11. Antoni J., "Fast computation of the kurtogram for the detection of transient faults", Mechanical Systems and Signal Processing, 2007, 21(1): 108–124.
12. Amar M., Gondal I., and Wilson C., "Vibration Spectrum Imaging: A Novel Bearing Fault Classification Approach", IEEE Transactions on Industrial Electronics, 2015, 62: 494–502.
13. Yang Y., Yu D.J. and Cheng J.S., "A roller bearing fault diagnosis method based on EMD energy entropy and ANN", Journal of Sound and Vibration, 2006, 294: 269–277.

14. Cui L., Wang J. and Lee S., "Matching pursuit of an adaptive impulse dictionary for bearing fault diagnosis", Journal of Sound and Vibration, 2014, 333(10): 2840–2862.
15. Guo L., Gao H., Li J., Huang H. and Zhang X., "Machinery vibration signal denoising based on learned dictionary and sparse representation", Journal of Physics: Conference Series. IOP Publishing, 2015, 628(1): 012124.
16. Roweis S.T. and Saul L.K., "Nonlinear dimensionality reduction by locally linear embedding", Science, 2000, 290(5500): 2323–2326.
17. Tenenbaum J.B., De Silva V. and Langford J.C., "A global geometric framework for nonlinear dimensionality reduction", Science, 2000, 290(5500): 2319–2323.
18. Zhang Z. and Zha H., "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment", Journal of Shanghai University (English Edition), 2004, 8(4): 406–424.
19. Li M., Xu J., Yang J., Yang D. and Wang D., "Multiple manifolds analysis and its application to fault diagnosis", Mechanical Systems and Signal Processing, 2009, 23(8): 2500–2509.
20. He Q., Liu Y., Wang J. and Gong C., "Time-frequency manifold for gear fault signature analysis," Instrumentation and Measurement Technology Conference (I2MTC), 2011 IEEE. IEEE, 2011: 1–5.
21. He Q., Liu Y., Long Q., and Wang J., "Time-frequency manifold as a signature for machine health diagnosis," IEEE Transactions on Instrumentation and Measurement, 2012, 61(5): 1218–1230.
22. Wang J., He Q. and Kong F., "Automatic fault diagnosis of rotating machines by time-scale manifold ridge analysis", Mechanical Systems and Signal Processing, 2013, 40(1): 237–256.
23. He Q. and Wang X., "Time-frequency manifold correlation matching for periodic fault identification in rotating machines", Journal of Sound and Vibration, 2013, 332(10): 2611–2626.
24. Wang J. and He Q., "Exchanged ridge demodulation of time-scale manifold for enhanced fault diagnosis of rotating machinery", Journal of Sound and Vibration, 2014, 333(11): 2450–2464.
25. Wang X. and He Q., " Machinery Fault Signal Reconstruction Using Time-Frequency Manifold", Engineering Asset Management – Systems, Professional Practices and Certification, Tse, PW; Mathew, J; Wong, K; Lam, R; Ko, CN, Berlin: Springer-Verlag, Germany, 2015: 777–787.
26. He Q., Wang X. and Zhou Q., "Vibration sensor data denoising using a time-frequency manifold for machinery fault diagnosis", Sensors, 2013, 14(1): 382–402.
27. Kennel M.B., Brown R. and Abarbanel H.D.I., "Determining embedding dimension for phase-space reconstruction using a geometrical construction", Physical Review A, 1992, 45 (6): 3403.
28. Cao L., "Practical method for determining the minimum embedding dimension of a scalar time series", Physica D: Nonlinear Phenomena, 1997, 110(1): 43–50.
29. Takens F., "Detecting strange attractors in turbulence", Springer Berlin Heidelberg, 1981.
30. He, Q., Song H. and Ding X., "Sparse signal reconstruction based on time-frequency manifold for rolling element bearing fault signature enhancement", IEEE Transactions on Instrumentation and Measurement, 2016, 65(2): 482–491.
31. Bearing Data Center. Available online: http://csegroups.case.edu/bearingdatacenter/home (accessed on 5 March 2016).

Matching Demodulation Transform and Its Application in Machine Fault Diagnosis

Xuefeng Chen and Shibin Wang

Abstract In this chapter, matching demodulation transform (MDT), an iterative algorithm, is introduced to generate a time-frequency (TF) representation with satisfactory energy concentration, and thus to extract the highly oscillatory frequency-modulation (FM) feature of rotating machine fault. As opposed to conventional time-frequency analysis (TFA) methods, this algorithm does not have to devise ad hoc parametric TF dictionary. Assuming the FM law of a signal can be well characterized by a determined mathematical model with reasonable accuracy, the MDT algorithm can adopt a partial demodulation and stepwise refinement strategy for investigating TF properties of the signal. The practical implementation of the MDT involves an iterative procedure that gradually matches the true instantaneous frequency (IF) of the signal. Moreover, because the MDT is a linear TFA method, it can reconstruct individual components from a multicomponent signal's TF representation. Theoretical analysis of the MDT's performance is provided, including quantitative analysis of the IF estimation error and the convergence condition. The validity and practical utility of the MDT is then demonstrated on simulation study, an experiment rotor system and a practical heavy oil catalytic cracking machine set with rotor rub-impact fault. The analysis results show that the MDT method is powerful in the analysis of FM signals and is an effective tool for the feature extraction of machine faults.

1 Introduction

Rotating machinery plays an important role in industrial applications, and its fault diagnosis is useful to prevent economic loss and catastrophic failure. Because vibration signals carry key information related to the health condition of rotating machinery, the study concerning how to extract useful feature from vibration signals has been attracted considerable interests in recent years [1, 2]. Some effective

X. Chen (✉) · S. Wang
Xi'an Jiaotong University, Xi'an, People's Republic of China
e-mail: chenxf@mail.xjtu.edu.cn

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors, Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_7

methods have been widely studied and used, such as empirical mode decomposition and Hilbert-Huang transform (HHT) [3–6], wavelet transform [7–9], time-frequency analysis (TFA) based methods [10–13], and sparsity-based methods [14–17].

Rotor, as one of the key components of mechanical devices, is widely used in rotating mechanical equipment. Rub-impact is a common and serious fault in the rotor system. The rub-impact phenomenon occurs when a rotating element periodically hits a stationary part in rotating machinery. Causes of rubbing can be imbalances, thermal misalignment, rotor/stator relative motion, fluid-dynamic forces producing instabilities and self-excited vibrations [18]. Accordingly, the vibration signal collected from rotor system will present nonlinear FM feature. It has a periodic time-varying instantaneous frequency (IF). The extraction of the periodic oscillatory IF is one effective way to diagnose the rub-impact fault in rotor systems [5].

The concept of IF is a nature extension of the conventional Fourier frequency. It describes how fast a signal oscillates locally at a time instant, or more generally, the different rates of oscillation at a given time [19, 20]. Thus, for nonstationary signal, the frequency at a particular time is well depicted by the concept of IF, and in many practical applications, such as radar and sonar [21], communications [22], biomedical engineering [23], and mechanical engineering [11, 24], IF characterizes important physical information of the signal, although the concept of IF still remains somewhat heuristic and lacks a rigorous and satisfactory mathematical definition.

Diverse IF estimation methods have been proposed to analyze FM signals masked by noise [20]. Among these IF estimators, the phase differencing-based and TFA-based methods are two well-known classes. The former is based on the definition, given by Gabor and later Ville, that the IF of a real signal is the derivative of the phase of its analytic signal. The effectiveness of this method is greatly hampered by noise. TFA itself is the core of the latter, TFA-based IF estimation methods. TFA provides a powerful tool to effectively characterize the time-frequency (TF) pattern of nonstationary signals. TF representations obtained by TFA methods map a one-dimensional signal, as a function of time only, to a two-dimensional function of time and frequency, and therefore give insight into the complex structure of the signal consisting of several components [25–27]. Therefore, TFA-based IF estimation method is an effective way to extract the IF from FM signals. However, the effectiveness depends on the property of the time-frequency representation (TFR) about concentrating the energy of a signal at and around the IF in the TF plane. The first moment of the TFR and the local maxima of TFR are two kinds of TFA-based IF estimation methods. The first moment estimate provides an unbiased estimation and is not affected by the multiplicative noise. However, the presence of additive noise leads to the serious degradation. It may have a high statistical variance even at high values of input SNR. The local maxima estimate is based on the detection of distributed maxima, it is hence used for signals contaminated with the additive noise [28, 29].

There exist many types of TFA methods, and the majority can be divided into two categories: the linear and the quadratic transforms. In linear transforms, the signal is characterized by its inner product with a dictionary of TF atoms, generated from a basis function by translation, modulation or dilation operations. Many linear methods, including conventional non-parametric methods (such as short-time Fourier transform (STFT), wavelet transform (WT)) and parametric methods (such as chirplet transform, local polynomial Fourier transform [30], adaptive STFT [31], and generalized demodulation approach [32]), make it possible to reconstruct the whole signal or parts of the signal. There is no doubt that the interpretation of the TF representation calculated by all these linear methods is dependent on the dictionary used in the method. Thus the essential issue of TFA methods is how to devise TF atoms that will be adapted to describe the energy density of a signal in time and frequency domain simultaneously. When adequate parameters are selected and TF atoms well present the IF trajectory of the signal, parametric TF transforms will be much more effective in characterizing the TF patterns of FM signals (even nonlinear FM signals) by providing a TF representation with satisfactory energy concentration. However, because of the complex parameters of parametric TFA methods, the excessive computational cost limits its application.

In this chapter, matching demodulation transform (MDT) [33, 34] is introduced as an iterative algorithm to generate a TFR with satisfactory energy concentration for multicomponent FM signal and to analyze the nonlinear FM signal of rotor rub-impact fault. Assuming the IF law of a signal can be well characterized by a determined mathematical model with reasonable accuracy, the MDT algorithm can adopt a partial demodulation and stepwise refinement strategy for investigating TF properties of the signal. However, we constantly confront applications where IF is unknown or difficult to be determined prior, especially for machine fault diagnosis. Even if it meets the seeming inapplicability in such situation, the applications can be well solved by the MDT, which involves an iterative procedure: a low-accuracy IF, estimated from a low-concentration TF representation, is used to roughly demodulate the signal and thus enhance the energy concentration of TF representation; then a high-accuracy IF, estimated from the enhanced representation, is used to further demodulate the signal and thus enhance the concentration again. In each iteration of the MDT algorithm, the estimated parameters of IF function can be used to build a bivariate demodulation operator and demodulate the signal partially into a bivariate signal with less FM, thus the corresponding TF representation has more concentrated energy distribution than the previous representation. With the implementation of the iterative procedure, the MDT gradually matches the true IF of the signal. This is the reason why this algorithm is named as “matching demodulation transform”. Moreover, the MDT’s performance is analyzed theoretically, including quantitative analysis of the IF estimation error and the convergence condition.

This iterative algorithm is adaptive to match the nonlinear IF rule of the analyzed signal, which enables the MDT to be suitable to extract the feature of periodic oscillatory IF for rotor fault diagnosis. The effectiveness of the method is verified by an experiment which is performed on Bently RK-4 Rotor Kit with a rub-impact

fault. Moreover, the application in rotor fault diagnosis of a heavy oil catalytic cracking machine set further demonstrates the effectiveness of the MDT. The periodic IF oscillation is successfully extracted. The comparison study indicates that the MDT behaves better than HHT.

The rest of the chapter is organized as follows. Section 2 reviews the background of the FM signal and STFT. The detailed information of the MDT's demodulation procedure for both mono- and multi- component signal is described in Sect. 3. The MDT's performance is analyzed theoretically in Sect. 4, including quantitative analysis of the IF estimation error and the convergence condition, and further analyzed by simulation study in Sect. 5. Section 6 provides an experimental verification of the MDT, which is followed by the application in a practical machine set with rotor rub-impact fault in Sect. 7. The conclusions are drawn in Sect. 8.

2 Theoretical Background

In this research, we will focus on signals which can be modelled with sums of sinusoidal functions, i.e. the signal $x(t)$ can be represented by the following model:

$$x(t) = \sum_{k=1}^{K} x_k(t) \tag{1}$$

with $x_k(t) = A_k(t)\cos(\phi_k(t))$ and $A_k(t) > 0,\ \phi_k'(t) > 0$, where the amplitude $A_k(t)$ and phase $\phi_k(t)$ are defined in terms of the analytic signal $z_k(t)$ given by Hilbert transform as follows

$$z_k(t) = x_k(t) + \mathrm{i}\mathcal{H}[x_k(t)], \tag{2}$$

and the Hilbert transform of $x_k(t)$ is defined as

$$\mathcal{H}[x_k(t)] = \pi^{-1}\mathrm{P.V.}\int_{-\infty}^{+\infty} \frac{x_k(\tau)}{t-\tau}\mathrm{d}\tau \tag{3}$$

where P.V. means that the integral is taken in the sense of the Cauchy principal value. The construction of the analytic signal $z_k(t)$ permits the amplitude $A_k(t)$ and phase $\phi_k(t)$ to be uniquely defined as

$$A_k(t)\mathrm{e}^{\mathrm{i}\phi_k(t)} = z_k(t) \tag{4}$$

and the original signal is recovered by $x_k(t) = \Re\{z_k(t)\}$. The instantaneous angular frequency of the component is the first derivative of the phase $\omega_k(t) = \phi_k'(t)$. Typically, the changes of $A_k(t)$ and $\phi_k'(t)$ are much slower than the change of $\phi_k(t)$

itself, which means that locally (i.e. in a short time interval) the component $x_k(t)$ can be regarded as a harmonic signal with amplitude $A_k(t)$ and frequency $\phi_k'(t)$.

In this model, if $K = 1$, the signal can be referred to as mono-component signal; if $K \geq 2$, the signal is referred to as multicomponent signal. The model defined by (1)–(4) applies for multicomponent signals and allows the modelling of K time-varying frequency laws.

2.1 Short-Time Fourier Transform

A linear TF transform correlates the signal with a dictionary of TF atoms that are concentrated in time and frequency. Let a general dictionary be denoted by $D = \{\phi_\gamma\}_{\gamma\in\Gamma}$ where γ may be a multi-index parameter set. Suppose that $\phi_\gamma \in L^2(\mathbb{R})$, the corresponding linear TF transform of $x \in L^2(\mathbb{R})$ is defined as

$$\Phi_x(\gamma) = \langle x, \phi_\gamma \rangle = \int_{-\infty}^{+\infty} x(t)\overline{\phi_\gamma(t)}\,\mathrm{d}t \tag{5}$$

where $\overline{\phi_\gamma(t)}$ is the complex conjugate of $\phi_\gamma(t)$, and $<x, \phi_\gamma>$ denotes the inner product of $x, \phi_\gamma \in L^2(\mathbb{R})$.

A short-time Fourier atom is constructed with a window function $g(t)$ modulated by the frequency ξ and translated by time-shift u:

$$\phi_\gamma(t) = g_{u,\xi}(t) = g(t-u)\mathrm{e}^{\mathrm{i}\xi(t-u)}$$

The resulting STFT of $x \in L^2(\mathbb{R})$ is

$$S_x(u,\xi) = \langle x, g_{u,\xi} \rangle = \int_{-\infty}^{+\infty} x(t)g(t-u)\,\mathrm{e}^{-\mathrm{i}\xi(t-u)}\,\mathrm{d}t \tag{6}$$

Because the Gaussian window function has the minimal area of the Heisenberg box, it is usually used as the window. The parametric Gaussian function is obtained by scaling the Gaussian function $g(t)$ by the variance σ:

$$g_\sigma(t) = \frac{1}{\sqrt{\sigma}} g\left(\frac{t}{\sigma}\right) = (\pi\sigma^2)^{-1/4}\mathrm{e}^{\frac{-t^2}{2\sigma^2}}$$

and the resulting STFT is

$$S_x(u,\xi) = \int_{-\infty}^{+\infty} x(t) g_\sigma(t-u)\, \mathrm{e}^{-\mathrm{i}\xi(t-u)}\, \mathrm{d}t = \int_{-\infty}^{+\infty} x(t+u) g_\sigma(t)\, \mathrm{e}^{-\mathrm{i}\xi t}\, \mathrm{d}t \quad (7)$$

2.2 Performance Analysis of the IF Estimator

Consider noisy discrete-time observations

$$x(nT) = x_0(nT) + \varepsilon(nT) \quad (8)$$

where $x_0(nT)$ is a sampled version of the continuous signal $x_0(t) = A(t)\mathrm{e}^{\mathrm{i}\phi(t)}$ with T being a sampling interval, and $\varepsilon(nT)$ is a complex-valued white Gaussian noise with i.i.d. real and imaginary parts. Thus, $\Re\{\varepsilon(nT)\}$ and $\Im\{\varepsilon(nT)\}$ are $\mathcal{N}\left(0, \sigma_\varepsilon^2/2\right)$ and the total variance of the noise is equal to σ_ε^2.

According to the STFT definition of the continuous signal in (7), the STFT of the discrete sequence $x(nT)$ is defined as

$$S_x(t,\omega) = T \sum_{n=-\infty}^{+\infty} x(t+nT) g_\sigma(nT)\, \mathrm{e}^{-\mathrm{i}\omega nT} \quad (9)$$

The spectrogram, as a TF energy density representation of the signal, is defined as

$$P(t,\omega) = T^2 \sum_{n_1=-\infty}^{+\infty} \sum_{n_2=-\infty}^{+\infty} x(t+n_1T)\, \overline{x(t+n_2T)}\, g_\sigma(n_1T) g_\sigma(n_2T)\, \mathrm{e}^{\mathrm{i}\omega(n_2-n_1)T} \quad (10)$$

The value of instantaneous angular frequency $\omega(t) = \phi'(t)$ can be estimated in the TF plane as

$$\tilde{\omega}(t) = \arg\max_{\omega} P(t,\omega) \quad (11)$$

The estimation error, at a time instant t, is defined as

$$\Delta\tilde{\omega}(t) = \tilde{\omega}(t) - \omega(t) \quad (12)$$

and due to the presence of the white Gaussian noise, the estimation error $\Delta\tilde{\omega}(t)$ can be considered as a random variable as well, and characterized by its bias $\mathrm{Bias}\{\Delta\tilde{\omega}(t)\}$ and variance $\mathrm{Var}\{\Delta\tilde{\omega}(t)\}$.

Proposition 1 [35, 36] Let $\tilde{\omega}(t)$ be a solution of (11), and the continuous signal $x_0(t) = A(t)\,\mathrm{e}^{\mathrm{i}\phi(t)}$ satisfies: $A \in C^1(\mathbb{R})$, $\phi \in C^\infty(\mathbb{R})$, and $|A'(t)| \ll |\phi'(t)|$. As $T \to 0$, the bias and variance of the IF estimator are given by

$$\text{Bias}\{\Delta\tilde{\omega}(t)\} \to \sum_{k=1}^{+\infty} \frac{\phi^{(2k+1)}(t)\sigma^{2k} M_{1,2k+2}}{(2k+1)! M_{1,2}} \tag{13}$$

$$\text{Var}\{\Delta\tilde{\omega}(t)\} \to \frac{\sigma_\varepsilon^2 M_{2,2}}{2|A(t)|^2 (M_{1,2})^2} \left[1 + \frac{\sigma_\varepsilon^2 T M_{2,0}}{\sigma |A(t)|^2 (M_{1,0})^2}\right] \frac{T}{\sigma^3} \tag{14}$$

where $M_{m,k} = \int_{-\infty}^{+\infty} t^k g^m(t)\,\mathrm{d}t$.

Special case: For the Gaussian window function $g(t) = (\pi)^{-1/4} \mathrm{e}^{-t^2/2}$, we have $M_{1,0} = \sqrt{2}\pi^{1/4}$, $M_{2,0} = 1$, $M_{1,2} = \sqrt{2}\pi^{1/4}$, $M_{2,2} = 1/2$, and $\frac{M_{1,2k+2}}{(2k+1)!} = \frac{M_{1,2}}{2^k k!}$. Then

$$\text{Bias}\{\Delta\tilde{\omega}(t)\} \to \sum_{k=1}^{+\infty} \frac{\phi^{(2k+1)}(t)\sigma^{2k}}{2^k k!}$$

$$\text{Var}\{\Delta\tilde{\omega}(t)\} \to \frac{\sigma_\varepsilon^2}{8\sqrt{\pi}|A(t)|^2} \left(1 + \frac{\sigma_\varepsilon^2 T}{2\sqrt{\pi}\sigma |A(t)|^2}\right) \frac{T}{\sigma^3}$$

Remark The value of the estimation error is determined by the variance σ_ε^2 of noise, the variance σ of the window function, and the nonlinear degree of the signal IF law. (1) The increase of the variance σ_ε^2 of the noise induces the increase of the variance of the estimation error. (2) The value of variance σ affects the estimation error from two aspects: increasing the window size σ decreases the variance of the estimation error but increases its bias. These two aspects should be taken into account to achieve a trade-off. (3) The nonlinear degree of the signal IF law induces the estimation error.

Note that if the nonlinear degree of the IF law can be decreased to a lower level, then we can use a larger variance σ to reduce the variance of the estimation error, and the bias can be kept at a low value. Therefore, we can get a more accurate IF estimation and a more concentrated TF representation of the signal. In this research, we propose an iterative algorithm, called matching demodulation transform, using a partial demodulation and stepwise refinement strategy, to decrease the nonlinear degree of the IF law, and thus to get an accurate IF estimation and a concentrated TF representation of a signal.

3 Matching Demodulation Transform

Matching demodulation transform (MDT) is an iterative algorithm to gradually improve the TF representation of signals. In this section, we first motivate the idea of MDT for both mono- and multi-component FM signal with determined frequency-modulation (FM) sources, and then give an iterative procedure for practical implementation of MDT.

3.1 Matching Demodulation Transform for Mono-Component Signal

To motivate the idea of MDT, let us start with an analytic mono-component FM signal

$$z(t) = A(t)\mathrm{e}^{\mathrm{i}\phi(t)} = A(t)\mathrm{e}^{\mathrm{i}[2\pi f_c t + \varphi(t)]}, \tag{15}$$

with a carrying frequency f_c and a determined FM source $\varphi(t)$. The IF of the signal is

$$f_{i,z}(t) = \phi'(t)/2\pi = f_c + \varphi'(t)/2\pi. \tag{16}$$

Considering that the FM source $\varphi(t)$ can be expanded into a linear part $\varphi(u) + \varphi'(u)(t-u)$ and a remainder $\Delta\varphi_u(t)$ in the vicinity of u, i.e.,

$$\varphi(t) = \varphi(u) + \varphi'(u)(t-u) + \Delta\varphi_u(t), \tag{17}$$

if we restrict ourselves to a small window in time around u, of the type $[u-\Delta T, u+\Delta T]$, with $\Delta T \approx 2\pi/\phi'(u)$, then the signal can be approximated as

$$\begin{aligned} z(t)|_{[u-\Delta T, u+\Delta T]} &\approx A(u)\,\mathrm{e}^{\mathrm{i}[\phi(u)+\phi'(u)(t-u)]} \\ &= A(u)\,\mathrm{e}^{\mathrm{i}[2\pi f_c t + \varphi(u) + \varphi'(u)(t-u)]}, \end{aligned} \tag{18}$$

which is essentially a truncated Taylor expansion in which terms on the order $O(A'(u))$, $O(\varphi''(u))$, have been neglected.

According to the determined FM source $\varphi(t)$, a bivariate function of time variable t and time-shift variable u:

$$f_d(t,u) = \mathrm{e}^{-\mathrm{i}[\varphi(t)-\varphi(u)-\varphi'(u)(t-u)]} = \mathrm{e}^{-\mathrm{i}\Delta\varphi_u(t)} \tag{19}$$

is introduced as a demodulation operator to transform the original signal $z(t)$ into a bivariate signal

$$z_d(t,u) = z(t)\cdot f_d(t,u) = A(t)\,\mathrm{e}^{\mathrm{i}[2\pi f_c t + \varphi(u) + \varphi'(u)(t-u)]}. \tag{20}$$

Thus,

$$z(t)|_{[u-\Delta T,\, u+\Delta T]} \approx z_d(t,u). \tag{21}$$

The STFT of the bivariate signal $z_d(t,u)$ is defined as

$$S_{z_d}(u,\xi) = \int_{-\infty}^{+\infty} z_d(t,u) g_\sigma(t-u)\, \mathrm{e}^{-\mathrm{i}\xi(t-u)}\, \mathrm{d}t, \tag{22}$$

Then the TF representation of the signal $z(t)$ can be approximated by

$$S_z(u,\xi) \approx S_{z_d}(u,\xi). \tag{23}$$

Because the MDT is a linear TFA method, it can reconstruct the signal from its MDT representation by the inverse MDT of $S_{z_d}(u,\xi)$ defined as

$$z(t) = \frac{1}{2\pi \|g_\sigma(t)\|^2} \cdot \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} S_{z_d}(u,\xi) \cdot \overline{f_d(t,u)} \cdot g_\sigma(t-u)\, \mathrm{e}^{\mathrm{i}\xi(t-u)}\, \mathrm{d}u\mathrm{d}\xi \tag{24}$$

In the demodulation procedure of this motivating example, the bivariate demodulation operator of time and time-shift is introduced to transform the one-dimensional signal, as a function of time only, into a two-dimensional bivariate function of time and time-shift. Moreover, the IF of the bivariate signal $z_d(t,u)$ is

$$f_{i,z_d}(t) = \frac{1}{2\pi}\frac{\partial}{\partial t}\{2\pi f_c t + \varphi(u) + \varphi'(u)(t-u)\} = f_c + \varphi'(u)/2\pi, \tag{25}$$

which is identical to the IF of the original signal at the corresponding time instant u, i.e.,.

$$f_{i,z_d}(t) = f_{i,z}(u), \tag{26}$$

In fact, the bivariate demodulation operator $f_d(t,u)$ eliminates the high order terms of the FM source $\varphi(t)$.

Figure 1 illustrates the demodulation procedure at the time instant u. To begin with, the spectrum of the windowed signal is widely spread around the frequency $f_c + \varphi'(u)/2\pi$. The demodulation operation includes two steps. Firstly, a forward

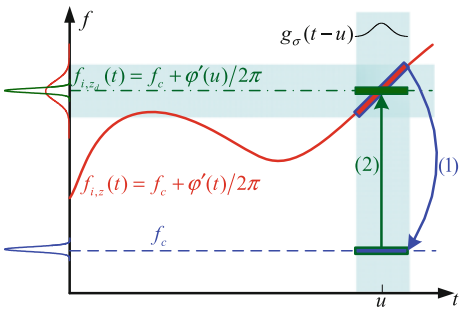


Fig. 1 The illustration of the MDT (*solid line* the IF law of the original nonlinear FM signal; *dash line* the IF law of the forward demodulated signal; *dash dot line* the IF law of the backward modulated signal $z_d(t,u)$)

demodulation operator $e^{-i\varphi(t)}$ which is a univariate function of the time variable t only, is used to demodulate the original signal into a pure carrier signal $A(t)\,e^{i2\pi f_c t}$, i.e., the nonlinear IF law is transformed into a constant frequency f_c. Thus, the spectrum of the demodulated signal is more concentrated than the original spectrum. However, this demodulated spectrum can not reflect the actual TF pattern of the original signal. Therefore, this demodulated signal is further modulated to another pure carrier signal by a backward modulation operator $e^{i[\varphi'(u)t+\varphi(u)-\varphi'(u)u]}$, a bivariate function of time variable t and time-shift variable u, and the frequency is equal to the IF of the original signal at the time instant u. Moreover, the concentration of the spectrum is preserved. As a whole, this two-step procedure performed at the time instant u greatly enhances the concentration of the spectrum.

If we apply this procedure to the signal at each time instant, we can retrieve the signal's TF representation with excellent energy concentration at every instant. The next step is to represent the demodulated signal in the TF plane, that is, the two-dimensional demodulated signal is transformed into another two-dimensional function of time and frequency. The result has a concentrated TF representation and can well characterize TF properties of the signal.

3.2 Matching Demodulation Transform for Multicomponent Signal

In some real-life applications, analyzed signals can be characterized as multicomponent signals, with TF representation consisting of several parallel (or approximately parallel) linear or nonlinear energy paths with low instantaneous bandwidth. For example, the Doppler signature of micro-multipath signals in over-the-horizon radar is typically composed of three components having close nonlinear TF behaviors [37, 38]. In these cases, each component $z_k(t)$ of the signal has the same modulation source $\varphi(t)$ as follows

$$z(t)=\sum_{k=1}^{K} z_k(t)=\sum_{k=1}^{K} A_k(t)\,e^{i[2\pi f_{c,k}t+\varphi(t)]}. \tag{27}$$

Because of the same modulation source, if one component can be demodulated by an appropriate demodulation operator, the other components would be demodulated by the same operator.

However, in many practical applications, the analyzed signal can be characterized as multicomponent signals with different modulation sources,

$$z(t)=\sum_{k=1}^{K} z_k(t)=\sum_{k=1}^{K} A_k(t)\,e^{i[2\pi f_{c,k}t+\varphi_k(t)]}. \tag{28}$$

The IF of the kth component is

$$f_{i,z_k}(t) = f_{c,k} + \varphi_k'(t) / 2\pi. \tag{29}$$

In this research, we consider that the multicomponent signal is a superposition of K well-separated FM components with separation distance d, that is, the IF satisfies

$$f_{i,z_k}(t) - f_{i,z_{k-1}}(t) \geq d, \ \forall t \in \mathbb{R}.$$

In this situation, the MDT runs into an inevitable problem that a single demodulation operator cannot be equally suitable for demodulating the mixed multiple components subjected to distinct modulation sources. So each component should have its own demodulation operator. The signal is demodulated to different forms by these demodulation operators, and their different TF representations concentrate around their corresponding IF trajectories in the TF plane. The following work is to fuse these different TF representations. Because all components are well separated, and each component has its own TF subregion, in this research, the fusion strategy is to partition the TF plane according to the TF patterns of the signal. That is, according to the IF trajectory of each component, the corresponding TF subregion is picked, divided from the TF plane, so that the information contained can be processed individually, and eventually all processed TF representations in different subregions are assembled together to provide the TF representation of the analyzed signal.

The TF lower boundary and TF upper boundary of the TF subregion of the component z_k are

$$\omega_{lb,k}(t) = \pi f_{i,z_k}(t) + \pi f_{i,z_{k-1}}(t), \quad \text{for } 1 < k \leq K, \tag{30}$$

$$\omega_{ub,k}(t) = \pi f_{i,z_k}(t) + \pi f_{i,z_{k+1}}(t), \quad \text{for } 1 \leq k < K, \tag{31}$$

where $\omega_{lb,1}(t)$ is set to zero and $\omega_{ub,K}(t)$ equals to the sample frequency of the signal system. Then, each component has its own TF subregion:

$$D_k = \left\{(t, \omega) : \omega_{lb,k}(t) \leq \omega < \omega_{ub,k}(t)\right\}. \tag{32}$$

According to the MDT algorithm for monocomponent signals, the bivariate demodulation operator for the component z_k is

$$f_{d,k}(t,u) = \mathrm{e}^{-\mathrm{i}\left[\varphi_k(t) - \varphi_k(u) - \varphi_k'(u)(t-u)\right]} = \mathrm{e}^{-\mathrm{i}\Delta\varphi_{k,u}(t)}. \tag{33}$$

This operator can be used to demodulate the component $z_k(t)$ into the corresponding carrier signal,

$$z_{d,k,k}(t,u) = z_k(t) \cdot f_{d,k}(t,u) = A_k(t)\, \mathrm{e}^{\mathrm{i}\left[2\pi f_{c,k} t + \varphi_k(u) + \varphi_k'(u)(t-u)\right]}, \tag{34}$$

and its IF is

$$f_{i,z_{d,k,k}}(t) = f_{c,k} + \varphi'_k(u) / 2\pi. \tag{35}$$

Then, the analytic signal $z(t)$ is transformed into

$$z_{d,k}(t,u) = z_{d,k,k}(t,u) + \sum_{j=1,j\neq k}^{K} z_j(t) \cdot f_{d,k}(t,u) = z_{d,k,k}(t,u) + \sum_{j=1,j\neq k}^{K} z_{d,j,k}(t,u) \tag{36}$$

where the component $z_{d,j,k}(t,u)$ is transformed from the jth component $z_j(t)$ via the kth bivariate demodulation operator $f_{d,k}(t,u)$, i.e.,

$$z_{d,j,k}(t,u) = z_j(t) \cdot f_{d,k}(t,u) = A_j(t)\, \mathrm{e}^{\mathrm{i}[2\pi f_{c,j}t + \varphi_j(t) - \Delta\varphi_{k,u}(t)]}, \tag{37}$$

and the IF is

$$f_{i,z_{d,j,k}}(t) = f_{c,j} + \left[\varphi'_j(t) - \varphi'_k(t) + \varphi'_k(u)\right] / 2\pi \quad \text{for} \quad j = 1, 2, \cdots, K, j \neq k.$$

When we restrict ourselves to a small window in time around u, of the type $[u - \Delta T,\, u + \Delta T]$, the IF can be approximated as

$$f_{i,z_{d,j,k}}(t)\big|_{[u-\Delta T,\, u+\Delta T]} \approx f_{c,j} + \varphi'_j(u) / 2\pi \approx f_{i,z_j}(t)\big|_{[u-\Delta T,\, u+\Delta T]}, \quad \text{for} \quad j = 1, 2, \cdots, K, j \neq k.$$

According to (36), the transformed signal via the kth bivariate demodulation operator consists of two parts: the pure carrier component $z_{d,k,k}(t,u)$ in (34) and the FM components $z_{d,j,k}(t,u)$, $j \neq k$ in (37). Then, the TF representation of the signal $z_{d,k}(t,u)$ also consists of two parts,

$$\begin{aligned} S_{z_{d,k}}(u,\xi) &= \int_{-\infty}^{+\infty} z_{d,k}(t,u) g_\sigma(t-u) \mathrm{e}^{-\mathrm{i}\xi(t-u)} \mathrm{d}t \\ &= \int_{-\infty}^{+\infty} z_{d,k,k}(t,u) g_\sigma(t-u) \mathrm{e}^{-\mathrm{i}\xi(t-u)} \mathrm{d}t \\ &\quad + \sum_{j=1,j\neq k}^{K} \int_{-\infty}^{+\infty} z_{d,j,k}(t,u) g_\sigma(t-u) \mathrm{e}^{-\mathrm{i}\xi(t-u)} \mathrm{d}t \\ &= S_{z_{d,k,k}}(u,\xi) + \sum_{j=1,j\neq k}^{K} S_{z_{d,j,k}}(u,\xi) \end{aligned} \tag{38}$$

Note that the TF representation of the pure carrier component $z_{d,k,k}(t,u)$ is well located around the IF trajectory $f_{i,z_k}(t)$ in the TF subregion D_k, and the others $z_{d,j,k}(t,u)$, $j \neq k$ are located around their respective IF trajectory $f_{i,z_j}(t)$ in their respective subregion D_j. Therefore, the TF representation of the analytic component $z_k(t)$ can be approximated by $S_{z_{d,k}}(u,\xi)$ in the TF subregion D_k,

$$S_{z_k}(u,\xi) \approx S_{z_{d,k,k}}(u,\xi) \approx \begin{cases} S_{z_{d,k}}(u,\xi), & (u,\xi) \in D_k \\ 0, & (u,\xi) \notin D_k. \end{cases}$$

Thus, the final TF fusion result of the signal $z(t)$ is the sum of all TF representations of K components,

$$S_z(u,\xi) = \sum_{k=1}^{K} S_{z_k}(u,\xi). \tag{39}$$

Meanwhile, the analytic component $z_k(t)$ can be reconstructed by the inverse MDT of $S_{z_k}(u,\xi)$ in the TF subregion

$$\begin{aligned} z_k(t) &= \frac{1}{2\pi\|g_\sigma(t)\|^2} \cdot \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} S_{z_{d,k,k}}(u,\xi) \cdot \overline{f_{d,k}(t,u)} \cdot g_\sigma(t-u) \mathrm{e}^{\mathrm{i}\xi(t-u)} \,\mathrm{d}u \mathrm{d}\xi \\ &\approx \frac{1}{2\pi\|g_\sigma(t)\|^2} \cdot \iint_{(u,\xi)\in D_k} S_{z_k}(u,\xi) \cdot \overline{f_{d,k}(t,u)} \cdot g_\sigma(t-u) \mathrm{e}^{\mathrm{i}\xi(t-u)} \,\mathrm{d}u \,\mathrm{d}\xi \\ &= \frac{1}{2\pi\|g_\sigma(t)\|^2} \cdot \iint_{(u,\xi)\in D_k} S_z(u,\xi) \cdot \overline{f_{d,k}(t,u)} \cdot g_\sigma(t-u) \mathrm{e}^{\mathrm{i}\xi(t-u)} \,\mathrm{d}u \,\mathrm{d}\xi \end{aligned} \tag{40}$$

Thus, the analytic signal $z(t)$ can be reconstructed as the sum of all the reconstructed components:

$$\begin{aligned} z(t) &= \sum_{k=1}^{K} z_k(t) \\ &\approx \sum_{k=1}^{K} \left[\frac{1}{2\pi\|g_\sigma(t)\|^2} \cdot \iint_{(u,\xi)\in D_k} S_z(u,\xi) \cdot \overline{f_{d,k}(t,u)} \cdot g_\sigma(t-u) \, \mathrm{e}^{\mathrm{i}\xi(t-u)} \,\mathrm{d}u \,\mathrm{d}\xi \right] \end{aligned} \tag{41}$$

Note that the MDT for multicomponent signal is an improvement of the MDT for monocomponent signal. Each component should be associated with an individual demodulation operator and an individual TF subregion so that the sum of all representations in different subregions can be used to illustrate the actual TF pattern of multicomponent signal.

3.3 Practical Iterative Implementation of Matching Demodulation Transform

In the previous two subsections, the MDT algorithm assumes that the FM law of a signal can be well characterized by a determined model, and thus a bivariate operator can be adopted to demodulate the signal and investigate TF properties of the signal. However, in most applications, the IF or the modulation source of the analyzed FM signal is unknown. The direct demodulation strategy is therefore unavailable. The practical implementation involves a novel iterative procedure: a low-accuracy IF estimated from a low-concentration TF representation is used to roughly demodulate the FM signal and enhance the energy concentration of the TF representation; then a high-accuracy IF estimated from the enhanced representation is used to further demodulate the signal and thus enhance the concentration again. With the implementation of the iterative procedure, the MDT gradually matches the true IF of the signal. This is the reason why this algorithm is named as "matching demodulation transform".

The pseudocode for the MDT's iterative implementation is shown in Fig. 2. The iterative procedure matches the true IF of the analyzed signal step by step, and the corresponding demodulation operator is used to demodulate the signal and enhance the concentration of its TF representation. In this subsection, some specifics of the implementation are described in detailed.

Algorithm The MDT iterative implementation
Initialization:
Choose parameters
the initial variance $\sigma_{(0)}$ of the window function,
the threshold δ,
the maximum iteration L.
Choose the IF model $\tilde{f}_i$.
STFT and the initial IF estimation:
Calculate the STFT S_z of the signal $z(t)$.
Estimate the initial IF $\tilde{f}_{i,(1)}$ based on S_z.
Repeat
Update the variance $\sigma_{(m)}$;
Calculate the TF representation $S_{z,(m)}$ by the MDT with $\tilde{f}_{i,(m)}$;
Estimate the IF $\tilde{f}_{i,(m+1)}$ based on $S_{z,(m)}$;
Calculate the terminating criterion MSE;
Update the iterative number $m \leftarrow m+1$.
Until $m > L$ or terminal condition is satisfied
Return $S_{z,(m)}$ and $\tilde{f}_{i,(m+1)}$

Fig. 2 Pseudocode for the MDT iterative implementation

(1) *The initialization of the MDT algorithm:* In the initialization stage, we have to choose the parameters of the MDT, including the variance σ of the window function and parameters for iterative terminal condition. Moreover, we have to choose the IF model to fit the discrete TF ridges for IF estimation.
As we known, the TF resolution is determined by the variance σ of the window function, which further affects the estimation error from two aspects, as mentioned in Sect. 2. In this paper, the initial variance $\sigma_{(0)}$ is approximately inversely proportional to the analyzed frequency band.
On the other hand, the IF model should be chosen to fit all the extracted discrete TF ridges. If the mathematical model with undetermined parameters is available in some application, we can use this model to estimate the IF of the FM signal, and the parameters of this model can be identified using least square method (LSM). In cases where the IF function is unknown, two general models can be used as expedient alternatives, which also turn out to derive satisfactory results.

- The polynomials: $f_i(t) = \sum_{k=0}^{K} \alpha_k t^k,\ \alpha_k \in \mathbb{R}$. Mathematically, the Weierstrass approximation theorem guarantees that any continuous function on a closed and bounded interval can be uniformly approximated on that interval by polynomials to any degree of accuracy. The polynomials has been used to approximate the IF law in many researches [11, 30]. In the polynomials model, the parameter α_0 can be considered as carrying frequency, and the other part is set as $m(t)$, i.e.

$$m(t) = \sum_{k=1}^{K} \alpha_k t^k. \tag{42}$$

Thus, the FM source $\varphi(t)$ is

$$\varphi(t) = 2\pi \int_0^t m(u)\,\mathrm{d}u = 2\pi \sum_{k=1}^{K} \frac{\alpha_k}{k+1} t^{k+1}. \tag{43}$$

Then the forward demodulation operator is

$$f_d^{\mathrm{I}}(t,u) = \mathrm{e}^{-\mathrm{i}\varphi(t)} = \mathrm{e}^{-\mathrm{i}2\pi \sum_{k=1}^{K} \frac{\alpha_k}{k+1} t^{k+1}}, \tag{44}$$

the backward modulation operator is

$$f_d^{\mathrm{II}}(t,u) = \mathrm{e}^{\mathrm{i}\left[\varphi'(u)(t-u)+\varphi(u)\right]} = \mathrm{e}^{\mathrm{i}2\pi \sum_{k=1}^{K} \left[\alpha_k u^k t + \left(\frac{\alpha_k}{k+1} - \alpha_k\right) u^{k+1}\right]}, \tag{45}$$

and the bivariate demodulation operator is

$$\begin{aligned} f_d(t,u) &= \mathrm{e}^{-\mathrm{i}[\varphi(t)-\varphi(u)-\varphi'(u)(t-u)]} \\ &= \mathrm{e}^{-\mathrm{i}2\pi\sum_{k=1}^{K}\left[\frac{\alpha_k}{k+1}\left(t^{k+1}-u^{k+1}\right)-\alpha_k u^k(t-u)\right]} \end{aligned} \tag{46}$$

- The Fourier series: $f_i(t) = \alpha_0 + \sum_{k=1}^{K}[\alpha_k \cos(\omega_k t) + \beta_k \sin(\omega_k t)]$. Fourier series is extremely useful as a way to break up an arbitrary periodic function into a set of simple terms that can be plugged in, solved individually, and then recombined to obtain the solution to the original problem or approximation to it to whatever accuracy is desired or practical. The Fourier series has been used to approximate the IF law, such as generalized warblet transform [39]. In this model, the parameter α_0 can be considered as carrying frequency, and the other part is set as $m(t)$, i.e.

$$m(t) = \sum_{k=1}^{K}[\alpha_k \cos(\omega_k t) + \beta_k \sin(\omega_k t)]. \tag{47}$$

Thus, the FM source $\varphi(t)$ is

$$\varphi(t) = 2\pi\int_0^t m(u)\,\mathrm{d}u = 2\pi\sum_{k=1}^{K}\left[\frac{\alpha_k}{\omega_k}\sin(\omega_k t) - \frac{\beta_k}{\omega_k}\cos(\omega_k t)\right]. \tag{48}$$

Then the forward demodulation operator is

$$\begin{aligned} f_d^{\mathrm{I}}(t,u) &= \mathrm{e}^{-\mathrm{i}\varphi(t)} \\ &= \mathrm{e}^{-\mathrm{i}2\pi\sum_{k=1}^{K}\left[\frac{\alpha_k}{\omega_k}\sin(\omega_k t)-\frac{\beta_k}{\omega_k}\cos(\omega_k t)\right]}, \end{aligned} \tag{49}$$

the backward modulation operator is

$$\begin{aligned} f_d^{\mathrm{II}}(t,u) &= \mathrm{e}^{\mathrm{i}[\varphi'(u)t+\varphi(u)-\varphi'(u)u]} \\ &= \mathrm{e}^{\mathrm{i}2\pi\sum_{k=1}^{K}\left[(\alpha_k\cos(\omega_k u)+\beta_k\sin(\omega_k u))(t-u)+\left(\frac{\alpha_k}{\omega_k}\sin(\omega_k u)-\frac{\beta_k}{\omega_k}\cos(\omega_k u)\right)\right]}, \end{aligned} \tag{50}$$

and the bivariate demodulation operator is

$$f_d(t,u) = \mathrm{e}^{-\mathrm{i}\left[\varphi(t)-\varphi(u)-\varphi'(u)(t-u)\right]} \cdot \mathrm{e}^{-\mathrm{i}2\pi\sum_{k=1}^{K}\left[\begin{array}{c}\left(\frac{\alpha_k}{\omega_k}\sin(\omega_k t)-\frac{\beta_k}{\omega_k}\cos(\omega_k t)\right)-\left(\frac{\alpha_k}{\omega_k}\sin(\omega_k u)-\frac{\beta_k}{\omega_k}\cos(\omega_k u)\right) \\ -(\alpha_k\cos(\omega_k u)+\beta_k\sin(\omega_k u))(t-u)\end{array}\right]} \tag{51}$$

(2) *IF estimation:* To guarantee the robustness against noise, the IF estimation is of paramount importance to be considered during the iterative procedure. In this paper, the IF estimation method is based on the local maximum energy in the TF representation in (11). Because of the existence of noise in practical applications, there are more than one maximum at some time even for mono-component signals. Thus the practical strategy for local maxima detection is based on an energy threshold, which is associated with the global maximum energy in the TF plane, to determine whether a local maximum is an IF ridge or not. Moreover, to improve the robustness against noise, the energy threshold should be closer to the global maximum with increased noise.
When the signal contains multiple well-separated components, the local maxima can detect the evolution of each component's IF over time. Thus we can use the distance d to separate different components. After the local maxima detection and separation, each IF can be estimated by least-square fitting, and their corresponding subregion D_k can be calculated as $D_k = \{(t,\omega) : \omega_{lb,k}(t) \le \omega < \omega_{ub,k}(t)\}$.

(3) *Terminal condition:* The iterative procedure terminates until no more evident change can be detected between two successive estimated IFs, and for each component, the termination condition is expressed as

$$MSE = \frac{\left\|\tilde{f}_{i,(m+1)}(t) - \tilde{f}_{i,(m)}(t)\right\|_2}{\left\|\tilde{f}_{i,(m)}(t)\right\|_2} < \delta_M \tag{52}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm, $\tilde{f}_{i,(m)}(t)$ denotes the estimated IF from the (m-1)th iteration and δ_M is a predetermined threshold.

(4) *The variance σ update:* Note that the value of the variance σ of Gaussian function similarly affects the error of IF estimator by MDT from two aspects: increasing the variance σ decreases the variance of the estimation error, but increases its bias. Normally, these two aspects should be taken into account to achieve a trade-off. Since the MDT gradually matches the true IF of the signal with the implementation of the iterative procedure, and the bias decreases accordingly, the parameter σ can increase to reduce the variance of the IF estimation error to achieve a new trade-off. That is to say, a small value of σ can be used first to determine the rough IF estimation; when the nonlinear degree of the IF law of the signal is decreased, a larger value can be used to obtain a high frequency resolution suitable for a robust IF estimation. Moreover, in a weak

noisy situation, the parameter σ can increase in a fast way for rapid convergence, for example,

$$\sigma_{(m)} = (m+1)\sigma_{(0)}, \tag{53}$$

where $\sigma_{(m)}$ denotes the variance σ in the mth iteration and $\sigma_{(0)}$ denotes the initial variance. While in a strong noisy situation, the parameter σ increase in a slow way for robust IF estimation, for example,

$$\sigma_{(m)} = \log_2(m+1)\sigma_{(0)}. \tag{54}$$

4 Performance Analysis of Matching Demodulation Transform

We provide a theoretical analysis of the MDT's performance in this section, including quantitative analysis of IF estimation error, and convergence condition and discussion.

4.1 Quantitative Analysis of IF Estimation Error

Let $r_{(m)}(t)$ denote the unknown phase function of the original signal $z(t) = A(t)\,\mathrm{e}^{\mathrm{i}\phi(t)}$ in the mth iteration, which is estimated from the $(m-1)$th iterative result of MDT, and $r_{(1)}(t)$ denote the phase function estimated from the spectrogram. The corresponding demodulation operator is $f_{d,(m)}(t,u) = \mathrm{e}^{-\mathrm{i}\Delta r_{(m),u}(t)}$, where the phase $\Delta r_{(m),u}(t)$ is a remainder of the first-order Taylor expansion of the estimated phase $r_{(m)}(t)$ in the vicinity of u, i.e.,

$$\Delta r_{(m),u}(t) = r_{(m)}(t) - [r_{(m)}(u) + r'_{(m)}(u)(t-u)] = \sum_{k=2}^{\infty} r_{(m)}^{(k)}(u)\frac{(t-u)^k}{k!} \tag{55}$$

The original signal is demodulated into a bivariate signal

$$z_d(t,u) = A(t)\,\mathrm{e}^{\mathrm{i}[\phi(t)-\Delta r_{(m),u}(t)]}. \tag{56}$$

According to the same noisy discrete-time observations in (8), the bivariate signal in the discrete form is

$$z_d(nT,u) = [A(nT)\,\mathrm{e}^{\mathrm{i}\phi(nT)} + \varepsilon(nT)] \cdot \mathrm{e}^{-\mathrm{i}\Delta r_{(m),u}(nT)}.$$

According to the MDT definition of the continuous signal in (22), the MDT of the discrete sequence $z(nT)$ is the STFT of the demodulated signal $z_d(nT,u)$, defined as

$$S_{z_d}(u,\xi) = T\sum_{n=-\infty}^{+\infty} z_d(u+nT,u)g_\sigma(nT)\,\mathrm{e}^{-\mathrm{i}\xi nT}, \tag{57}$$

where

$$z_d(u+nT,u) = [A(u+nT)\,\mathrm{e}^{\mathrm{i}\phi(u+nT)} + \varepsilon(u+nT)] \cdot \mathrm{e}^{-\mathrm{i}\Delta r_{(m),u}(u+nT)}. \tag{58}$$

Similarly, a TF energy density representation of the signal can be defined by MDT as

$$\begin{aligned} P_{(m)}(t,\omega) = T^2 \sum_{n_1=-\infty}^{+\infty}\sum_{n_2=-\infty}^{+\infty} z_d(t+n_1T,t)\overline{z_d(t+n_2T,t)}g_\sigma \\ (n_1T)g_\sigma(n_2T)\mathrm{e}^{\mathrm{i}\omega(n_2-n_1)T} \end{aligned} \tag{59}$$

Then, the value of instantaneous angular frequency can be estimated in the TF plane as

$$\tilde{\omega}_{(m)}(t) = \arg\max_{\omega} P_{(m)}(t,\omega), \tag{60}$$

Accordingly, the estimation error is

$$\Delta\tilde{\omega}_{(m)}(t) = \tilde{\omega}_{(m)}(t) - \omega(t), \tag{61}$$

and the estimation error $\Delta\tilde{\omega}_{(m)}(t)$ can be also considered to be a random variable, and characterized by its bias and variance.

Proposition 2 Let $\tilde{\omega}_{(m)}(t)$ be a solution of (60). The continuous signal $x_0(t) = A(t)\mathrm{e}^{\mathrm{i}\phi(t)}$ satisfies: $A \in C^1(\mathbb{R})$, $\phi \in C^\infty(\mathbb{R})$, and $|A'(t)| \ll |\phi'(t)|$, and the demodulation operator $f_d(t,u) = \mathrm{e}^{-\mathrm{i}\Delta r_{(m),u}(t)}$ satisfies: $r_{(m)} \in C^\infty(\mathbb{R})$. As $T \to 0$, the bias and variance of the IF estimator are given by

$$\mathrm{Bias}\{\Delta\tilde{\omega}_{(m)}(t)\} \to \sum_{k=1}^{+\infty} \frac{\left[\phi^{(2k+1)}(t) - r_{(m)}^{(2k+1)}(t)\right]\sigma^{2k}M_{1,2k+2}}{(2k+1)!M_{1,2}} \tag{62}$$

$$\mathrm{Var}\{\Delta\tilde{\omega}_{(m)}(t)\} \to \frac{\sigma_\varepsilon^2 M_{2,2}}{2|A(t)|^2 (M_{1,2})^2}\left(1+\frac{\sigma_\varepsilon^2 T M_{2,0}}{\sigma|A(t)|^2 (M_{1,0})^2}\right)\frac{T}{\sigma^3} \tag{63}$$

Proof Consider the Taylor expansion of the phase functions $\phi(t+nT)$, and the signal having slow varying amplitude, the bivariate signal in (59) is expanded as

$$\begin{aligned} z_d(t+nT,t) &= [A(t)\,\mathrm{e}^{\mathrm{i}[\phi(t)+\phi'(t)nT+\Delta\phi(t,nT)]} \\ &\quad +\varepsilon(t+nT)]\cdot \mathrm{e}^{-\mathrm{i}\Delta r_{(m),t}(t+nT)} \\ &= A(t)\,\mathrm{e}^{\mathrm{i}[\phi(t)+\phi\prime(t)nT+\Delta\phi(t,nT)-\Delta r_{(m),t}(t+nT)]} \\ &\quad +\varepsilon(t+nT)\cdot \mathrm{e}^{-\mathrm{i}\Delta r_{(m),t}(t+nT)} \end{aligned} \tag{64}$$

with $\Delta\phi(t,nT)=\sum_{k=2}^{+\infty}\phi^{(k)}(t)(nT)^k/k!$ and $\Delta r_{(m),t}(t+nT)=\sum_{k=2}^{+\infty} r_{(m)}^{(k)}(t)(nT)^k/k!$.

Since the IF is located at the stationary points of $P_{(m)}(t,\omega)$ which is determined by the zero value of the derivation of $P_{(m)}(t,\omega)$, that is, the IF estimate $\tilde{\omega}_{(m)}(t)$ is given by solving $\partial P_{(m)}(t,\omega)/\partial\omega=0$ for ω, where

$$\begin{aligned} \frac{\partial P_{(m)}(t,\omega)}{\partial\omega} &= T^2\sum_{n_1=-\infty}^{+\infty}\sum_{n_2=-\infty}^{+\infty} z_d(t+n_1T,t)\overline{z_d(t+n_2T,t)} \\ &\quad g_\sigma(n_1T)g_\sigma(n_2T)\mathrm{e}^{\mathrm{i}\omega(n_2-n_1)T}(\mathrm{i}(n_2-n_1)T) \end{aligned} \tag{65}$$

Any error in the IF estimate may be due to the one (or a combination) of the following [40]:

(1) the estimate error $\Delta\tilde{\omega}_{(m)}(t)$;
(2) the error due to the residual of the deviation $\delta_{\Delta\phi,\Delta r}$;
(3) the error due to the noise δ_ε.

Thus, the linearization of $\partial P_{(m)}(t,\omega)/\partial\omega=0$ with respect to these quantities gives

$$\begin{aligned} \frac{\partial P_{(m)}(t,\omega)}{\partial\omega} &= \left.\frac{\partial P_{(m)}(t,\omega)}{\partial\omega}\right|_0 + \left.\frac{\partial^2 P_{(m)}(t,\omega)}{\partial\omega^2}\right|_0 \Delta\tilde{\omega}_{(m)}(t) \\ &\quad +\left.\frac{\partial P_{(m)}(t,\omega)}{\partial\omega}\right|_0 \delta_{\Delta\phi,\Delta r} + \left.\frac{\partial P_{(m)}(t,\omega)}{\partial\omega}\right|_0 \delta_\varepsilon = 0 \end{aligned} \tag{66}$$

where $|_0$ means that the corresponding expressions are computed at the point $\omega=\phi'(t)$, $\Delta\phi(t,nT)=0$, $\Delta r_{(m),t}(t+nT)=0$ and $\varepsilon(nT)=0$. The term

$\partial P_{(m)}(t,\omega)/\partial\omega\big|_0\delta_{\Delta\phi,\Delta r}$ represents variation of the derivative $\partial P_{(m)}(t,\omega)/\partial\omega$ caused by small $\Delta\phi(t,nT)$ and $\Delta r_{(m),t}(t+nT)$, and the term $\partial P_{(m)}(t,\omega)/\partial\omega\big|_0\delta_\varepsilon$ caused by noise $\varepsilon(nT)$. Therefore, the general expression of the estimation error is

$$\Delta\tilde{\omega}_{(m)}(t) = -\left(\left.\frac{\partial P_{(m)}(t,\omega)}{\partial\omega}\right|_0+\left.\frac{\partial P_{(m)}(t,\omega)}{\partial\omega}\right|_0\delta_{\Delta\phi,\Delta r}+\left.\frac{\partial P_{(m)}(t,\omega)}{\partial\omega}\right|_0\delta_\varepsilon\right)\Big/\left(\left.\frac{\partial^2 P_{(m)}(t,\omega)}{\partial\omega^2}\right|_0\right) \tag{67}$$

and the elements of (67) are

$$\left.\frac{\partial P_{(m)}(t,\omega)}{\partial\omega}\right|_0 = 0$$

$$\left.\frac{\partial^2 P_{(m)}(t,\omega)}{\partial\omega^2}\right|_0 = -2|A(t)|^2T^2\sum_{n_1=-\infty}^{+\infty}\sum_{n_2=-\infty}^{+\infty}g_\sigma(n_1T)g_\sigma(n_2T)(n_1T)^2$$

$$\left.\frac{\partial P_{(m)}(t,\omega)}{\partial\omega}\right|_0\delta_{\Delta\phi,\Delta r} = |A(t)|^2T^2\sum_{n_1=-\infty}^{+\infty}\sum_{n_2=-\infty}^{+\infty}g_\sigma(n_1T)g_\sigma(n_2T)(\mathrm{i}(n_2-n_1)T)$$
$$\mathrm{e}^{\mathrm{i}\left[\left(\Delta\phi(t,n_1T)-\Delta r_{(m),t}(t+n_1T)\right)-\left(\Delta\phi(t,{}_2T)-\Delta r_{(m),t}(t+n_2T)\right)\right]}$$

$$\left.\frac{\partial P_{(m)}(t,\omega)}{\partial\omega}\right|_0\delta_\varepsilon = T^2\sum_{n_1=-\infty}^{+\infty}\sum_{n_2=-\infty}^{+\infty}g_\sigma(n_1T)g_\sigma(n_2T)\left[A(t)\mathrm{e}^{\mathrm{i}[\phi(t)+\phi'(t)n_1T]}+\varepsilon(t+n_1T)\right]$$
$$\times\left[\overline{A(t)\,\mathrm{e}^{\mathrm{i}[\phi(t)+\phi'(t)n_2T]}+\varepsilon(t+n_2T)}\right]\cdot\mathrm{e}^{\mathrm{i}\omega(n_2-n_1)T}(\mathrm{i}(n_2-n_1)T)$$

where the fact that $g_\sigma(t)$ is a real symmetric window function is considered. The only random term is $\partial P_{(m)}(t,\omega)/\partial\omega\big|_0\delta_\varepsilon$. The expression (67) then can be rewritten as

$$\Delta\tilde{\omega}_{(m)}(t) = \frac{|A(t)|^2L_{(m),\sigma}(t)+\Xi_\sigma}{2|A(t)|^2M_{T,1,2}M_{T,1,0}} \tag{68}$$

where the notation

$$
\begin{aligned}
L_{(m),\sigma}(t) &= T^2 \sum_{n_1=-\infty}^{+\infty} \sum_{n_2=-\infty}^{+\infty} g_\sigma(n_1T) g_\sigma(n_2T) (\mathrm{i}(n_2-n_1)T) \\
&\quad \mathrm{e}^{\mathrm{i}\left[\left(\Delta\phi(t,n_1T)-\Delta r_{(m),t}(t+n_1T)\right)-\left(\Delta\phi(t,n_2T)-\Delta r_{(m),t}(t+n_2T)\right)\right]} \\
&\approx T^2 \sum_{n_1=-\infty}^{+\infty} \sum_{n_2=-\infty}^{+\infty} g_\sigma(n_1T) g_\sigma(n_2T) ((n_1-n_2)T) \\
&\quad \left(\Delta\phi(t,n_1T) - \Delta r_{(m),t}(t+n_1T) - \Delta\phi(t,n_2T) + \Delta r_{(m),t}(t+n_2T)\right) \\
&= 2T^2 \sum_{k=1}^{+\infty} \left[\frac{\phi^{(2k+1)}(t) - r_{(m)}^{(2k+1)}(t)}{(2k+1)!} \sum_{n_1=-\infty}^{+\infty} g_\sigma(n_1T)(n_1T)^{2k+2} \right] \sum_{n_2=-\infty}^{+\infty} g_\sigma(n_2T) \\
\Xi_\sigma &= \left. \frac{\partial P_{(m)}(t,\omega)}{\partial \omega} \right|_0 \delta_\varepsilon \\
M_{T,m,k} &= T^m \sum_{n=-\infty}^{+\infty} g_\sigma^m(nT)(nT)^k
\end{aligned}
$$

are used. Considering that the expected value of Ξ_σ is equal to zero, i.e., $E\{\Xi_\sigma\} = 0$, thus the expected value of the estimation error $\Delta\tilde{\omega}_{(m)}(t)$, that is, the bias is

$$
\mathrm{Bias}\left\{\Delta\tilde{\omega}_{(m)}(t)\right\} = E\left\{\Delta\tilde{\omega}_{(m)}(t)\right\} = \frac{|A(t)|^2 L_{(m),\sigma}(t) + E\{\Xi_\sigma\}}{2|A(t)|^2 M_{T,1,2} M_{T,1,0}} = \frac{L_{(m),\sigma}(t)}{2M_{T,1,2}M_{T,1,0}} \tag{69}
$$

The variance of the IF estimation error $\Delta\tilde{\omega}_{(m)}(t)$ is

$$
\begin{aligned}
\mathrm{Var}\left\{\Delta\tilde{\omega}_{(m)}(t)\right\} &= \frac{\mathrm{Var}\{\Xi_\sigma\}}{4|A(t)|^4 \left(M_{T,1,2}M_{T,1,0}\right)^2} \\
&= \frac{|A(t)|^2 \sigma_\varepsilon^2 M_{T,2,2} \left(M_{T,1,0}\right)^2 + \sigma_\varepsilon^4 M_{T,2,2} M_{T,2,0}}{2|A(t)|^4 \left(M_{T,1,2}M_{T,1,0}\right)^2}
\end{aligned} \tag{70}
$$

For $T \to 0$, we have the following approximations:

$$M_{T,m,k} = T^m \sum_{n=-\infty}^{+\infty} g_\sigma^m(nT)(nT)^k \to \sigma^{k-m/2+1} T^{m-1} M_{m,k}$$

$$L_{(m),\sigma}(t) \approx 2T^2 \sum_{k=1}^{+\infty} \left[\frac{\phi^{(2k+1)}(t) - r_{(m)}^{(2k+1)}(t)}{(2k+1)!} \sum_{n_1=-\infty}^{+\infty} g_\sigma(n_1 T)(n_1 T)^{2k+2} \right] \sum_{n_2=-\infty}^{+\infty} g_\sigma(n_2 T)$$

$$\to 2 \sum_{k=1}^{+\infty} \frac{\left[\phi^{(2k+1)}(t) - r_{(m)}^{(2k+1)}(t)\right] \sigma^{2k+3}}{(2k+1)!} M_{1,2k+2} M_{1,0}$$

with $M_{m,k} = \int_{-\infty}^{+\infty} t^k g^m(t)\,\mathrm{d}t$. Then, the bias and the variance of the IF estimation error are

$$\mathrm{Bias}\left\{\Delta\tilde{\omega}_{(m)}(t)\right\} \to \sum_{k=1}^{+\infty} \frac{\left[\phi^{(2k+1)}(t) - r_{(m)}^{(2k+1)}(t)\right] \sigma^{2k} M_{1,2k+2}}{(2k+1)! M_{1,2}},$$

$$\mathrm{Var}\left\{\Delta\tilde{\omega}_{(m)}(t)\right\} \to \frac{|A(t)|^2 \sigma_\varepsilon^2 \sigma T M_{2,2} (M_{1,0})^2 + \sigma_\varepsilon^4 T^2 M_{2,2} M_{2,0}}{2|A(t)|^4 \sigma^4 (M_{1,2} M_{1,0})^2}$$

$$= \frac{\sigma_\varepsilon^2 M_{2,2}}{2|A(t)|^2 (M_{1,2})^2} \left(1 + \frac{\sigma_\varepsilon^2 T M_{2,0}}{\sigma |A(t)|^2 (M_{1,0})^2}\right) \frac{T}{\sigma^3}.$$

■

Remark The value of the estimation error in (62) and (63) is determined by the variance σ_ε^2 of noise, the variance σ of the window function, and the estimation accuracy in the $(m-1)$th MDT iteration. The former two factors are same with the effect in the Proposition 1 for STFT, and the last factor is the essential difference between the MDT-based IF estimation and the STFT-based IF estimation. The more accurate the IF estimation in the $(m-1)$th MDT iteration is, the smaller the error between the estimated IF $r'_{(m)}(t)$ and the actual IF $\phi'(t)$ is, and smaller the bias in the mth MDT iteration is. It means that the phase function of the bivariate demodulation operator can "weaken" the modulation of the signal, thus the IF estimated from the TF representation of the demodulated signal is more accurate than the one estimated from the TF representation of the original signal.

Proposition 3 If the phase function of the bivariate demodulation operator satisfies the condition:

$$\begin{aligned} 0 &< \sum_{n=-\infty}^{+\infty} g_\sigma(nT) nT \left[r_{(m)}(t+nT) - r'_{(m)}(t) nT \right] \\ &< \sum_{n=-\infty}^{+\infty} g_\sigma(nT) nT [\phi(t+nT) - \phi'(t) nT] \end{aligned} \tag{71}$$

or

$$\begin{aligned} 0 &> \sum_{n=-\infty}^{+\infty} g_\sigma(nT) nT \left[r_{(m)}(t+nT) - r'_{(m)}(t) nT \right] \\ &> \sum_{n=-\infty}^{+\infty} g_\sigma(nT) nT [\phi(t+nT) - \phi'(t) nT] \end{aligned} \tag{72}$$

we have

$$\left| \text{Bias}\{\Delta\tilde{\omega}_{(m)}(t)\} \right| < |\text{Bias}\{\Delta\tilde{\omega}(t)\}|.$$

Proof: Since

$$\begin{aligned} \phi(t+nT) &= \phi(t) + \phi'(t) nT + \Delta\phi(t, nT), \\ r_{(m)}(t+nT) &= r_{(m)}(t) + r'_{(m)}(t) nT + \Delta r_{(m),t}(t+nT), \end{aligned}$$

we have

$$\begin{aligned} L_{(m),\sigma}(t) &\approx 2T^2 \sum_{n_1=-\infty}^{+\infty} \left[g_\sigma(n_1 T) n_1 T \left(\Delta\phi(t, n_1 T) - \Delta r_{(m),t}(t+n_1 T) \right) \right] \\ &\quad \sum_{n_2=-\infty}^{+\infty} g_\sigma(n_2 T) \\ &= 2T^2 \sum_{n_1=-\infty}^{+\infty} \left[g_\sigma(n_1 T) n_1 T \left(\phi(t+n_1 T) - \phi'(t) n_1 T - r_{(m)}(t+n_1 T) + r'_{(m)}(t) n_1 T \right) \right] \\ &\quad \sum_{n_2=-\infty}^{+\infty} g_\sigma(n_2 T) \end{aligned}$$

where the fact that $g_\sigma(t)$ is a real symmetric window function is considered. If the condition (71) is satisfied, we have

$$\begin{aligned} 0 &< \sum_{n=-\infty}^{+\infty} \left[g_\sigma(nT) nT \left(\phi(t+nT) - \phi'(t) nT - r_{(m)}(t+nT) + r'_{(m)}(t) nT \right) \right] \\ &< \sum_{n=-\infty}^{+\infty} [g_\sigma(nT) nT (\phi(t+nT) - \phi'(t) nT)], \end{aligned}$$

so

$$\left|\frac{L_{(m),\sigma}(t)}{2M_{T,1,2}M_{T,1,0}}\right| < \left|\frac{L_{\sigma}(t)}{2M_{T,1,2}M_{T,1,0}}\right|,$$

where the notation

$$L_{\sigma}(t) \approx 2T^2 \sum_{n_1=-\infty}^{+\infty} [g_{\sigma}(n_1T)n_1T(\phi(t+n_1T) - \phi'(t)n_1T)] \sum_{n_2=-\infty}^{+\infty} g_{\sigma}(n_2T),$$

is used. Then

$$\left|\text{Bias}\left\{\Delta\tilde{\omega}_{(m)}(t)\right\}\right| \leq \left|\text{Bias}\{\Delta\tilde{\omega}(t)\}\right|.$$

The other condition is analogous. ■

Special case: For the aforementioned motivating example, the phase of the bivariate demodulation operator satisfies $r_{(m)}(t) = \phi(t)$. Thus we have $\phi^{(2k+1)}(t) = r_{(m)}^{(2k+1)}(t)$, for $k \in \mathbb{Z}^+$, and

$$\text{Bias}\left\{\Delta\tilde{\omega}_{(m)}(t)\right\} = 0.$$

Remark In this special case, the bivariate demodulation operator eliminates the estimation error due to the residual of the deviation $\delta_{\Delta\phi}$, which is an error in the IF estimation based on STFT. Despite a general demodulation operator, the bias of error for the MDT-based IF estimator is smaller than the one for STFT-based IF estimator if the condition (71) or (72) is satisfied, which means that the phase function of the bivariate demodulation operator can "weaken" the modulation of the signal, thus the IF estimated from the TF representation of the demodulated signal is more accurate than the one estimated from the TF representation of the original signal.

4.2 Convergence Condition and Discussion

We assume that the phase of the original signal can be modelled by the function $\phi(t)$ with some determined parameter $\boldsymbol{\alpha}$, i.e., the phase function being denoted by $\phi(t, \boldsymbol{\alpha})$, then the mth estimated phase function $r_{(m)}(t)$ can be denoted by $\phi(t, \boldsymbol{\alpha}_{(m)})$, i.e., $r_{(m)}(t) = \phi(t, \boldsymbol{\alpha}_{(m)})$. If the condition (71) or (72) is satisfied for $m = M$, then the bias of the IF estimation error $\Delta\tilde{\omega}_{(M)}(t)$, based on the TF representation of the mth MDT iteration, is smaller than the bias of $\Delta\tilde{\omega}(t)$ from the spectrogram, which

means that the estimated IF $r'_{(M+1)}(t) = \phi'(t, \boldsymbol{\alpha}_{(M+1)})$ is closer to the true IF $\phi'(t, \boldsymbol{\alpha})$ than the initial estimated IF $r'_{(1)}(t) = \phi'(t, \boldsymbol{\alpha}_{(1)})$ for $t \in \mathbb{R}$. That is to say, the estimated parameters $\boldsymbol{\alpha}_{(M+1)}$ of the IF model $\phi'(t, \boldsymbol{\alpha}_{(M+1)})$ is closer to the actual parameters $\boldsymbol{\alpha}$ of the true IF $\phi'(t, \boldsymbol{\alpha})$. Therefore, the condition (71) or (72) is further satisfied for $m = M + 1$, and the signal can be better demodulated by the operator associated with $\phi'(t, \boldsymbol{\alpha}_{(M+1)})$, so the bias of the corresponding IF estimation error $\Delta\tilde{\omega}_{(M+1)}(t)$ is further smaller than the bias of $\Delta\tilde{\omega}_{(M)}(t)$. In a word, the implementation of the iterative procedure will match the true IF step by step and finally converge to it.

On the other hand, from (62) and (63) it is clear that increasing the variance σ of the Gaussian function decreases the variance of the IF estimation error but increases its bias. These two aspects should be taken into account to achieve a bias-variance trade-off. The optical variance σ can be obtained by solving the following optimization problem (minimizing the mean squared error of the IF estimation) [40]:

$$\begin{aligned}\sigma_{\mathrm{opt}}(t) &= \arg\min_{\sigma} E\left\{\left(\Delta\tilde{\omega}_{(m)}(t)\right)^2\right\} \\ &= \arg\min_{\sigma}\left\{\mathrm{Var}\left\{\Delta\tilde{\omega}_{(m)}(t)\right\} + \left(E\left\{\Delta\tilde{\omega}_{(m)}(t)\right\}\right)^2\right\}.\end{aligned}$$

For the Gaussian window function with small σ, the optimal variance can be approximately expressed as

$$\sigma_{\mathrm{opt}}(t) \approx \left(\frac{3T\sigma_\varepsilon^2}{8\sqrt{\pi}|A(t)|^2}\left(1 + \frac{\sigma_\varepsilon^2}{2\sqrt{\pi}|A(t)|^2}\right) \Big/ \left(\phi^{(3)}(t) - r^{(3)}_{(m)}(t)\right)^2\right)^{1/7}.$$

This optimal variance depends on the third derivative of the phase function $\phi^{(3)}(t)$ and the second derivative of estimated IF $r^{(3)}_{(m)}(t)$, which is time and signal dependent. Note that if the third derivation $\phi^{(3)}(t)$ is significant different for different t, a time-varying variance of the window function is required for the optimization of the estimation accuracy. Without the effect of the demodulation $r_{(m)}(t)$, a time-invariant variance of window function is difficult to simultaneously minimize the IF estimation error for all the times.

If the condition (71) or (72) is satisfied, the MDT can match the true IF with the implementation of the iterative procedure, and the difference between two derivation $\phi^{(3)}(t)$ and $r^{(3)}_{(m)}(t)$ is reduced, thus the optimal variance σ can increased to achieve a new bias-variance trade-off. Moreover, because of the demodulation effect, the difference $\phi^{(3)}(t) - r^{(3)}_{(m)}(t)$ is much less than the third derivation $\phi^{(3)}(t)$ itself, thus a time-invariant variance of window function can minimize the IF estimation error simultaneously. The improvement in estimation accuracy is significant even without a time-varying variance.

5 Simulation Study

In this section we utilize a range of simulation examples to illustrate the effectiveness of the MDT. For all examples in this section, MDT is carried out starting from a Gaussian window function; other window functions that are well localized in frequency give similar results.

5.1 Applying the MDT to Simulation Signal

In this case, we consider a nonlinear FM signal, whose IF is an inverse hyperbolic sine function given by

$$f_i(t) = f_{c0} + a_0 \cdot \sinh^{-1}(b_0(t - c_0)),$$

where $\sinh^{-1}(\cdot)$ is an inverse hyperbolic sine function; the parameters are $f_{c0} = 256$, $a_0 = 40$, $b_0 = 100$ and $c_0 = 0.5$, respectively. This inverse hyperbolic sine function IF is shown in Fig. 3a. According to this IF, set $m_0(t) = a_0 \sinh^{-1}(b_0(t - c_0))$, and $p_0(t) = \ln\left(b_0(t - c_0) + \sqrt{b_0^2(t - c_0)^2 + 1}\right)$, then the modulation source is

$$\varphi_0(t) = 2\pi \int_{-\infty}^{t} m_0(u)\,\mathrm{d}u = 2\pi a_0\left[(t - c_0)p_0(t) - b_0^{-1}\cosh p_0(t)\right], \tag{73}$$

with $\cosh(\cdot)$ being a hyperbolic cosine function. Thus, the corresponding simulation signal is

$$x_0(t) = \cos\left(2\pi f_{c0}t + 2\pi a_0\left[(t - c_0)p_0(t) - b_0^{-1}\cosh p_0(t)\right]\right). \tag{74}$$

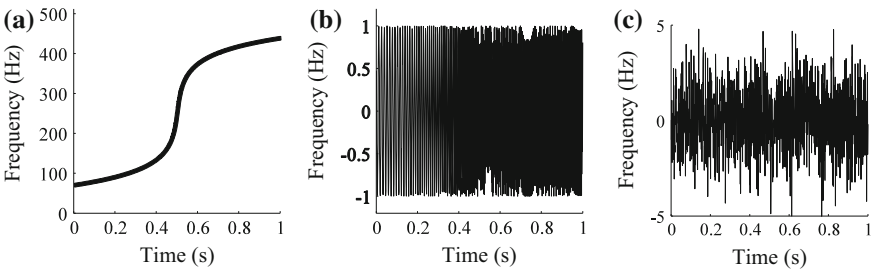


Fig. 3 The inverse hyperbolic sine IF and the simulation signal: **a** IF, **b** noise-free signal, and **c** noisy signal ($SNR = -6.65$ dB)

The waveform of this noise-free simulation signal is shown in Fig. 3b. The discrete signal has 1024 samples in the interval $t \in [0, 1]$.

In order to explore the tolerance to noise of the MDT algorithm, white Gaussian noise with variance $\sigma_\varepsilon^2 = 1.5$ is added to the simulation signal. Let $\varepsilon(t)$ denote the noise with zero mean and variance $\sigma_\varepsilon^2 = 1$. The noisy simulation signal is

$$x(t) = x_0(t) + 1.5\varepsilon(t). \tag{75}$$

The signal-to-noise ratio (SNR) is defined as

$$SNR = 10\log_{10}(P_x / P_n), \tag{76}$$

where P_x is the energy of the noiseless signal and P_n is the energy of the noise. The waveform of this noisy signal is shown in Fig. 3c. The SNR of this noisy simulation signal is $SNR = -6.65\,\text{dB}$.

The MDT method is applied to analyze the simulation signal. In this case, the IF model is the inverse hyperbolic sine function with unknown parameters, i.e., the IF model is

$$f_i(t) = f_c + a \cdot \sinh^{-1}(b(t-c)),$$

And the parameters f_c, a, b and c are unknown. Set $m(t) = a\sinh^{-1}(b(t-c))$, and $p(t) = \ln\left(b(t-c) + \sqrt{b^2(t-c)^2 + 1}\right)$, the modulation source is

$$\varphi(t) = \int_{-\infty}^{t} m(u)\,\mathrm{d}u = 2\pi a\left[(t-c)p(t) - b^{-1}\cosh p(t)\right], \tag{77}$$

Thus, the forward demodulation operator is

$$f_d^{\mathrm{I}}(t,u) = \mathrm{e}^{-\mathrm{i}\varphi(t)}, \tag{78}$$

the backward modulation operator is

$$f_d^{\mathrm{II}}(t,u) = \mathrm{e}^{\mathrm{i}[\varphi'(u)(t-u)+\varphi(u)]} = \mathrm{e}^{\mathrm{i}[m(u)(t-u)+\varphi(u)]}, \tag{79}$$

and the bivariate demodulation operator is

$$f_d(t,u) = \mathrm{e}^{-\mathrm{i}[\varphi(t)-\varphi(u)-\varphi'(u)(t-u)]} = \mathrm{e}^{-\mathrm{i}[\varphi(t)-\varphi(u)-m(u)(t-u)]}. \tag{80}$$

In this case, the initial variance of the window is $\sigma_{(0)} = 1/96$ and the threshold is $\delta = 10^{-4}$ for the MDT's MSE termination condition, the maximal number of iterations is 20. The STFT result and the MDT result are shown in Fig. 4. It can be found in the comparison that only a few part ridge in the STFT result can be observed and the most part are influenced by the noise. However, the IF of the

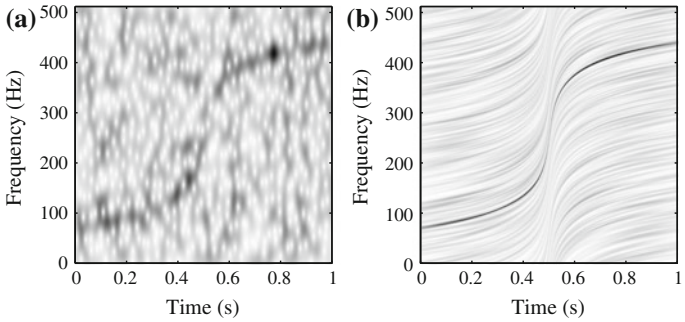


Fig. 4 **a** The STFT result and **b** the MDT result

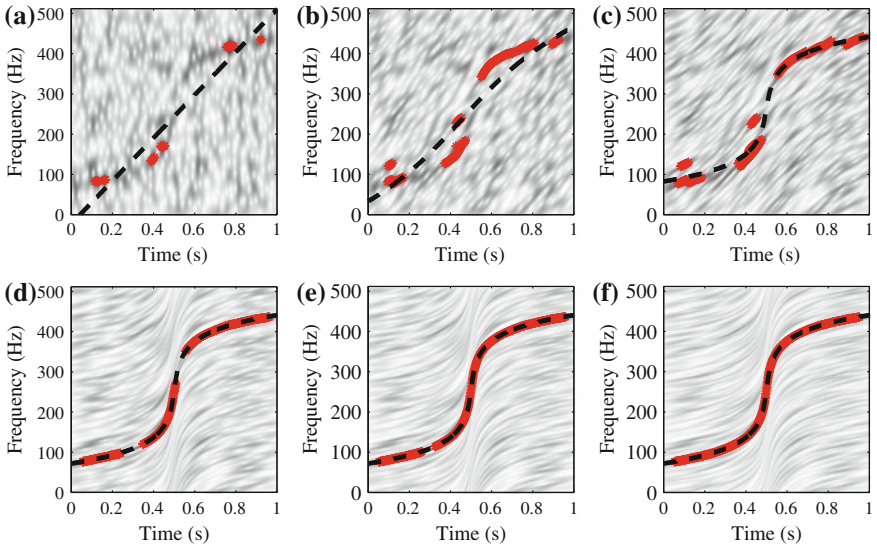


Fig. 5 The initial IF estimation based on the STFT representation and the first 5 iterations in the MDT iterative procedure

simulation signal is clearly represented in the MDT result with high energy concentration.

In order to illustrate the convergence procedure of the MDT method, Fig. 5 gives the initial IF estimation based on the STFT representation and the first 5 iterations in the MDT iterative procedure. In the TFR of each iteration, the dashed line describes the estimated IF by fitting the extracted TF ridges. It can be clearly observed that, with the iterative process, the estimated IF gradually convergent to the true IF of the simulation signal and the energy concentration of the TFR gradually is improved, even though the initial IF estimation is less accurate.

In order to further illustrate the convergence procedure of the MDT method, Fig. 6 shows the evolution of the logarithmic MSE while implementing the iterative procedure of the MDT algorithm in this noisy case, not only the MSE between the estimated IF and the simulated IF (denoted by MSE1 and marked with circles), but also the MSE between two successive estimated IFs (denoted by MSE2 and marked with squares). It can be observed that the accuracy of IF estimation is improved as the iteration proceeds on. Moreover, Fig. 7 provides the identified parameters in each MDT iteration, including f_c, a, b, and c. It also can be observed that the accuracy of identified parameters is improved as the iteration proceeds on. It thus illustrates the convergence discussion in Sect. 4.2.

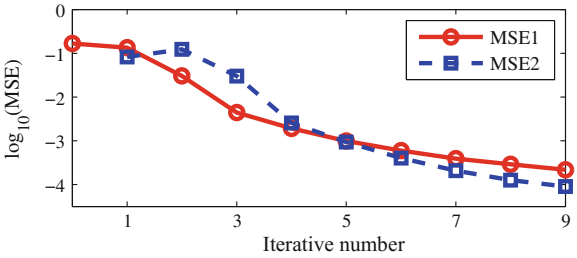


Fig. 6 The logarithmic MSE values of the iterative procedures (*circle* the MSE between the identified IF and the simulated IF; *square* the MSE between two successive identified IF)

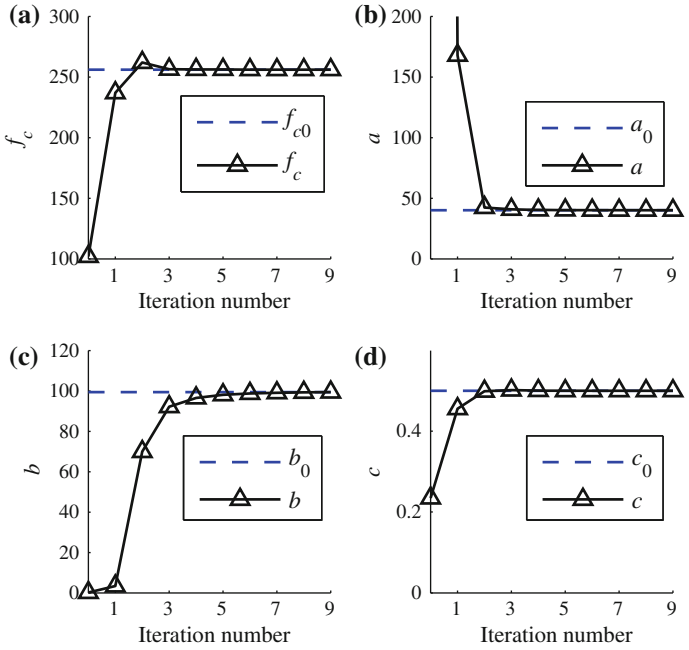


Fig. 7 Identified parameters in each iteration: **a** f_c, **b** a, **c** b, and **d** c

For comparison, some TFA methods are considered, including S-method [41, 42], Hilbert-Huang transform (HHT), reassigned STFT (RSTFT), reassigned smoothed pseudo Wigner-Ville distribution (RSPWVD), synchrosqueezed wavelet transform (SWT) [43] and generalized synchrosqueezing transform (GST) [44, 45]. The TF representations obtained by them are presented in Fig. 8. Moreover, the MSE values of the IF estimation based on these representations are listed in Table 1 to compare the performance of the MDT algorithm.

S-method belongs to the general class of smoothed pseudo Wigner-Ville distributions. Figure 8a shows the S-method result by applying a rectangular window with length 37. It can be seen that the representation has a low TF concentration. Because of the high frequency modulation, the S-method is incapable of tracking the rapid IF variation, and the strong FM part is cracked. Moreover, because of the influence of noise, the estimation error (MSE = 7.99×10^{-2}) of the IF tracking is much larger than the MDT result (MSE = 2.19×10^{-4}). The HHT is an extension of the EMD algorithm, which is the Hilbert energy spectrum of the decomposition result of the EMD. In this case, the HHT spectrum shown in Fig. 8b is incapable of characterizing the true nonlinear IF law. Figure 8c, d illustrate two reassignment methods, RSTFT and RSPWVD. The frequency smoothing window of the RSTFT is a Gaussian window with length 71. The time and frequency smoothing window of RSPWVD are the hamming window with length 103 and 257, respectively. The concentration of the RSTFT and RSPWVD results is improved, however the strong FM part is also cracked and the IF estimation error is larger than the MDT result. Figure 8e, f provide the results of the original SWT and an improved SWT, i.e., the GST. The Q-factors of wavelets used in SWT and GST are 20 and 30π,

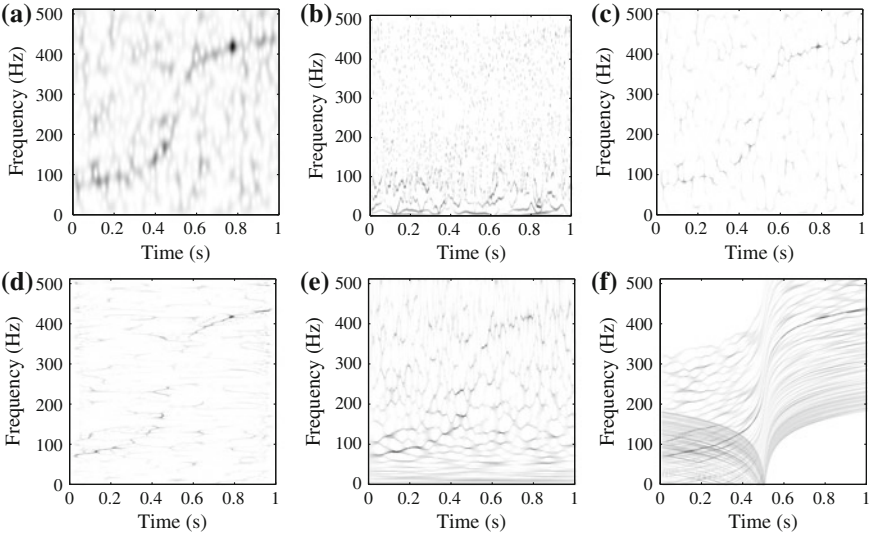


Fig. 8 The comparison study of the simulated signal. The TF representations generated by **a** S-method, **b** HHT, **c** RSTFT, **d** RSPWVD, **e** SWT, and **f** GST

Table 1 The MSE comparison for different methods

	MSE
MDT	2.19×10^{-4}
S-method	7.99×10^{-2}
RSTFT	7.52×10^{-2}
RSPWVD	5.56×10^{-2}
SWT	1.25×10^{-2}
GST	3.48×10^{-4}

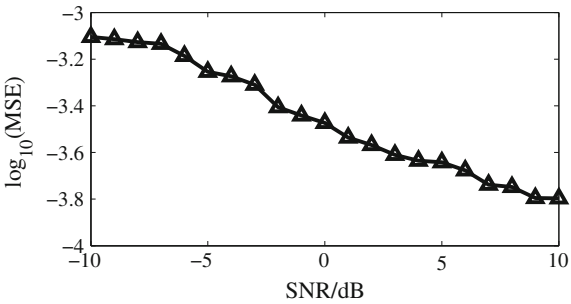


Fig. 9 The logarithmic MSE values of IF estimation by MDT for different noise levels

respectively. The SWT result is influenced by the noise and the nonlinear IF law. Although the concentration is essentially improved, the IF estimation error is still larger than the MDT result. The GST requires the phase function to demodulate the signal, thus the identified phase function by MDT is used in this paper. However, it can be seen from Table 1 that the IF estimation accuracy (MSE = 3.48×10^{-4}) of GST is lower than the accuracy of the used phase function identified by MDT (MSE = 2.19×10^{-4}). In comparison with these representations, the MDT algorithm provides TF representations with satisfying energy concentration and IF estimation method with reasonable accuracy. The nonlinear IF law can be clearly identified in the TF representation shown in Fig. 4b.

To further explore the tolerance to noise of the MDT, different noise levels are investigated in this paper. Figure 9 illustrates the MSE values of IF estimation while MDT implementing 10 iterations at different noise levels. It can be seen that the MSE of the IF estimation increases with noise. That is to say, the noise decreases the accuracy of the IF estimation. Despite the high noise level (−10 dB), the MDT algorithm converges to the true IF of the analyzed signal, and the MSE is 7.94×10^{-4}.

Finally, we apply the MDT algorithm to a multicomponent signal with three close IF signatures spaced 10 Hz apart. Each IF is an inverse hyperbolic sine function given by

$$f_{i,k}(t) = f_{c,k} + a_0 \cdot \sinh^{-1}(b_0(t - c_0)) \quad \text{for} \quad k = 1, 2, 3,$$

where the parameters are $f_{c,1} = 246$, $f_{c,2} = 256$, $f_{c,3} = 266$, $a_0 = 40$, $b_0 = 100$ and $c_0 = 0.5$, respectively. These IFs are shown in Fig. 10a. According to these Ifs, the signal is

$$x_k(t) = \cos\left(2\pi f_{c,k} t + 2\pi a_0 \left[(t-c)p_0(t) - b_0^{-1}\cosh p_0(t)\right]\right) \quad \text{for} \quad k = 1, 2, 3,$$

with $p_0(t) = \ln\left(b_0(t-c_0) + \sqrt{b_0^2(t-c_0)^2 + 1}\right)$. Then, the multicomponent signal is

$$x_0(t) = \sum_{k=1}^{3} x_k(t)$$

The waveform of this noise-free simulation signal is shown in Fig. 10b. The discrete signal has 1024 samples in the interval $t \in [0, 1]$. In order to explore the tolerance to noise of the MDT algorithm, white Gaussian noise is added to the simulation signal. The noisy simulation signal is

$$x(t) = x_0(t) + 1.5\,\varepsilon(t). \tag{81}$$

The SNR is −2.06 dB, and the waveform of this noisy signal is shown in Fig. 10c.

In this case, the initial variance of the window is $\sigma_{(0)} = 1/96$ and the threshold is $\delta = 2 \times 10^{-4}$ for the MDT's MSE termination condition, the maximal number of iterations is 20. The STFT result and the MDT result are shown in Fig. 11. It can be found that the three components in the STFT result are mixed with each other and can not be distinguished. However, the three IFs of the simulation multicomponent signal are clearly represented in the MDT result with high energy concentration.

For comparison, some TFA methods are considered, including S-method, RSTFT, and SWT. The TF representations obtained by them are presented in Fig. 12a–c, respectively. The length of the rectangular window used in the S-method is 17. The frequency smoothing window of the RSTFT is same as the window function used in MDT. The Q-factors of wavelets used in SWT and GST are also 20 and 30π, respectively. The former three TFA methods cannot distinguish the three components in the simulated signal. Compared with these representations, the MDT method provides TF representations with satisfying energy

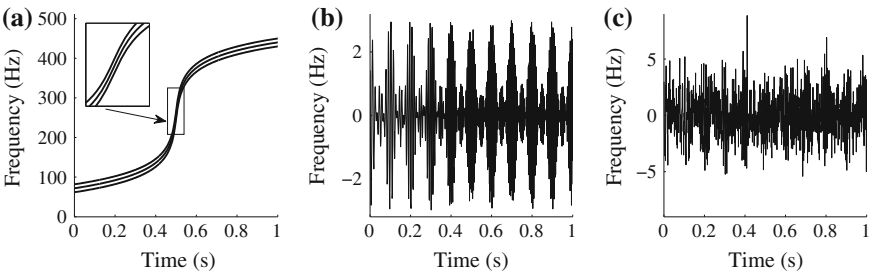


Fig. 10 The inverse hyperbolic sine IF and the simulation signal: **a** IF, **b** noise-free signal, and **c** noisy signal ($SNR = -2.06\,\text{dB}$)

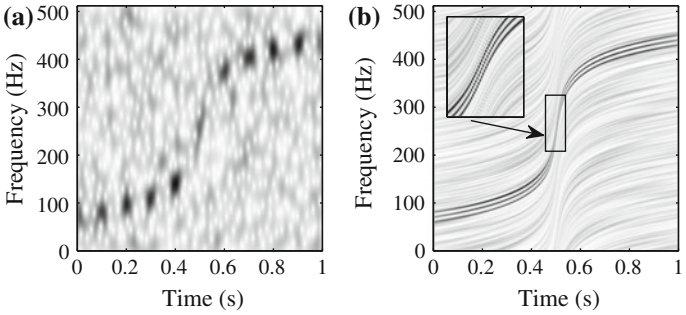


Fig. 11 **a** The STFT result and **b** the MDT result

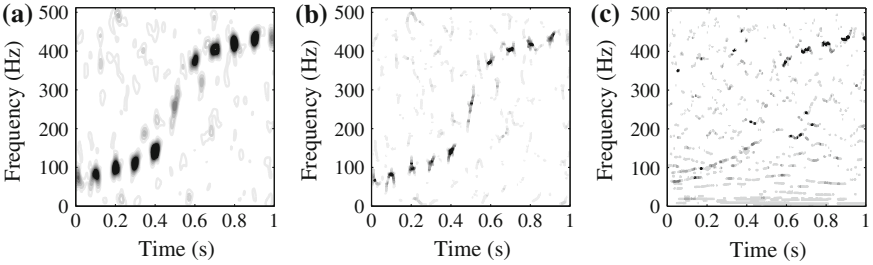


Fig. 12 The comparison study of the multicomponent signal: The TFR generated by **a** S-method, **b** RSTFT, and **c** SWT

concentration and the three components can be clearly distinguished and identified in Fig. 11b.

5.2 Applying Signal Reconstruction to Simulation Signal

In this subsection, we explore the reconstruction by the inverse MDT algorithm. As illustrated in the preceding section, the MDT representation has well-localized zones of concentration. Thus, one can use these to select the zone corresponding to one component, and then integrate, in the reconstruction formula (24), over this zone of the integration domain.

We firstly illustrate this with the monocomponent FM signal shown in Fig. 3. Figure 13 shows the example of the reconstruction, including the zone selection in the TF plane and the reconstruction result. The zone selected in this example is centred around the estimated IF trajectory on the MDT representation, which has a fixed width inversely proportional to the variance σ of the window. In order to reduce the influence of the noise, the width is as narrow as possible. Moreover, the selected zone should contain the component of the analyzed signal as much as

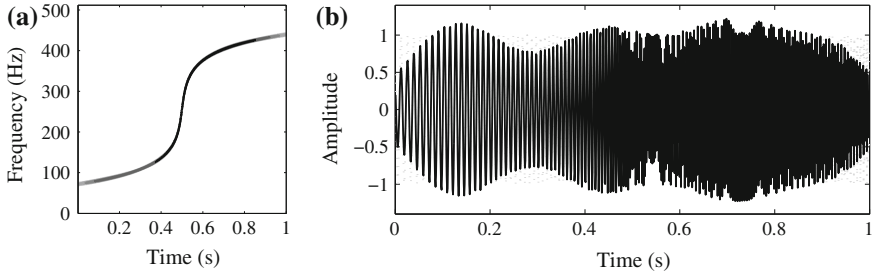


Fig. 13 The reconstruction of the inverse hyperbolic sine signal by the inverse MDT algorithm: **a** Zone selection for the reconstruction of the simulated signal, and **b** the reconstruction result according to the inverse MDT algorithm (*plotted* in *solid line* over the original simulated signal in *dot line*)

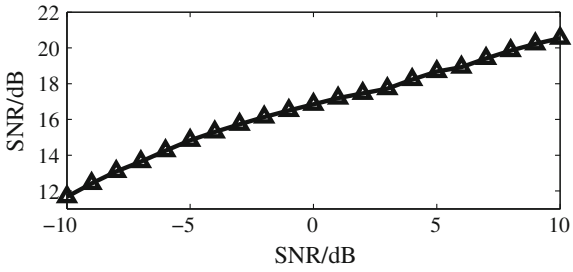


Fig. 14 The SNR values of reconstructed signals by the inverse MDT algorithm for different levels of noise

possible. These two aspects should be taken into account to achieve a trade-off. The SNR of the reconstructed signal shown in Fig. 13 is 13.98 dB. For the noisy signal, the added white Gaussian noise is widely spread out in the TF plane, so the noise will also spread out in the area of IF trajectory of the analyzed signal. Because the window used in the MDT is wide, the noise is smoothed and it is reconstructed into the signal, which leads to the amplitude error in the reconstructed signal.

To further explore the reconstruction of the inverse MDT algorithm, different noise levels are investigated in this paper. Figure 14 shows the SNR values of the reconstructed signals at different noise levels. It can be seen that the SNR values of reconstructed signals decrease linearly with increased noise.

Finally, we illustrate the signal reconstruction for individual components from the multicomponent FM signal shown in Fig. 10. Figure 15 shows the reconstructed three components. The SNR values of the reconstructed individual components are 12.37, 14.39 and 12.09 dB, respectively.

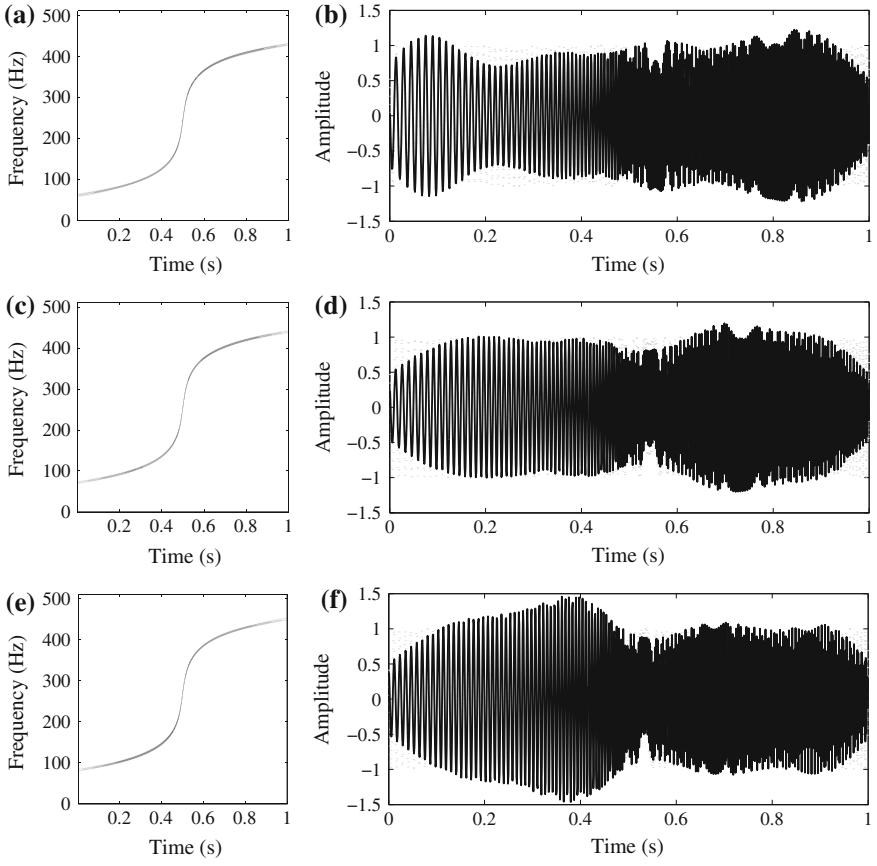


Fig. 15 The reconstruction of the multicomponent signal by the inverse MDT algorithm: **a** zone selection for the component $x_1(t)$, **b** the reconstruction result $\tilde{x}_1(t)$, **c** zone selection for the component $x_2(t)$, **d** the reconstruction result $\tilde{x}_2(t)$, **e** zone selection for the component $x_3(t)$, and **f** the reconstruction result $\tilde{x}_3(t)$

6 Experimental Verification

In this section we utilize a range of simulation examples to illustrate the effectiveness of the MDT. For all examples in this section, MDT is carried out starting from a Gaussian window function; other window functions that are well localized in frequency give similar results.

Rotor is one of the most important components in the rotating machinery. Rub-impact is a common nonlinear fault in a rotor system, which may bring a serious hazard to machines. Therefore, it is necessary to monitor the condition of the rotor system.

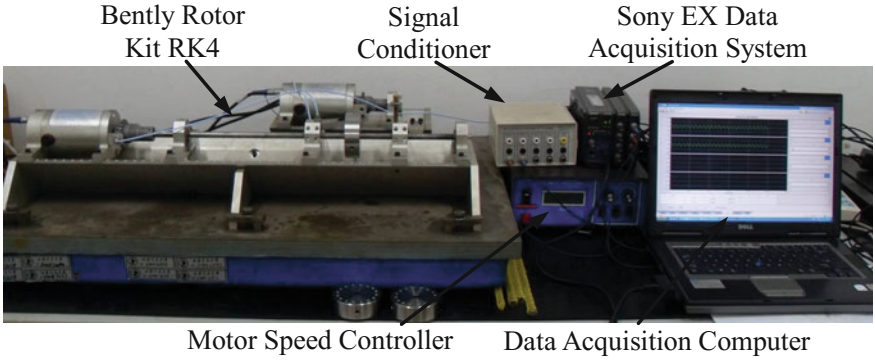


Fig. 16 The experiment set of Bently RK-4 rotor kit

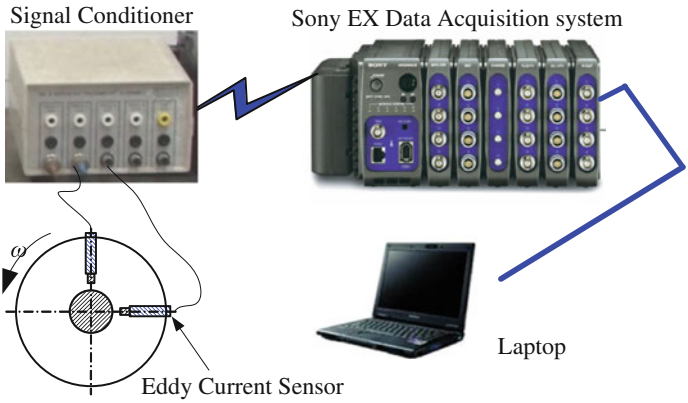


Fig. 17 The sketch of data acquisition system

To verify the effectiveness of the MDT in feature extraction for rotor fault diagnosis, a simulating experiment of rotor rub-impact fault was performed on Bently RK-4 Rotor Kit. Figure 16 shows an overall view of the experiment system, which includes Bently RK-4 Rotor Kit test rig, signal conditioner, and signal acquisition system.

A radial rubbing fault is simulated by using a rub screw to hit the radial surface of the shaft. The rub screw is secured in the mounting block with a locknut and it is adjusted to obtain rub-impact fault in different degrees. In this experiment, the rubbing fault is slight. Rotor displacement signals in the horizontal and the vertical directions are collected by eddy current sensors, which are mounted on the probe base. The sketch of data acquisition system is shown in Fig. 17. The operation speed of the rotor kit is 2000 r/min, the sample frequency is 2 kHz and 1024 points are sampled. The vibration signal and its spectrum are shown in Fig. 18. Because

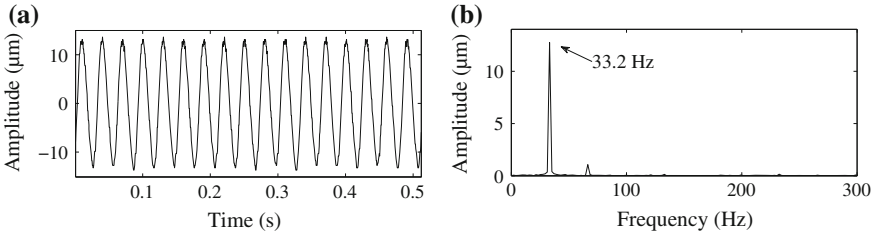


Fig. 18 **a** The vibration signal of the Bently rotor system and **b** its Fourier spectrum

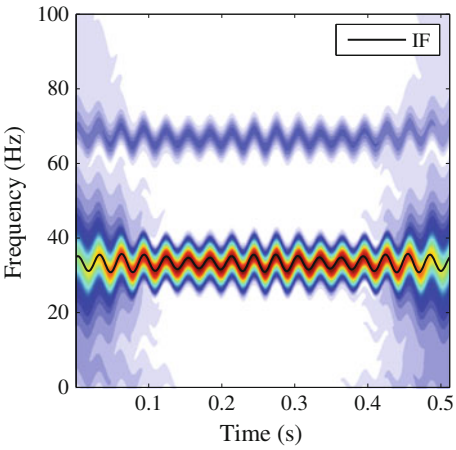


Fig. 19 The TFR of the vibration signal obtained by MDT and the extracted IF from this representation

the rubbing fault of this experiment is very slight, only the rotating frequency and its weak second harmonics can be observed in the spectrum of the vibration signal.

Because of the periodic rubbing between the rotating element and the stationary part, FM phenomenon will exist in the vibration signal, which is a periodic and nonlinear FM signal. It has a nonlinear time-varying IF. The MDT is applied to analyze the vibration signal and to investigate its property. The TFR result obtained by MDT and the extracted IF are shown in Fig. 19. In this case, the value of σ for the initial IF estimation is 0.005, and the threshold is $\delta = 10^{-3}$ for MDT's MSE termination condition. Figure 20 shows the evolution of the logarithmic MSE while implementing iterative procedure of the MDT algorithm. After five iterations, the algorithm converges to a periodic oscillated IF, which can be found in Fig. 19. That is to say, although the MDT is an iterative algorithm, and the calculation speed is an important factor for an iterative algorithm, in this case, the MDT uses only five iterations to converge to the periodic oscillated IF. Therefore, the calculation time of the MDT in this case is approximately five times the computing time of STFT. The extracted periodic oscillatory IF is the feature of periodic rubbing between the rotating element and the stationary part. Moreover, Fig. 21a redraws the extracted IF and its mean value which is approximately equal to the rotating frequency of the

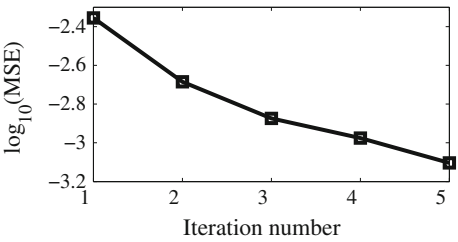


Fig. 20 The logarithmic MSE values of the iterative procedures

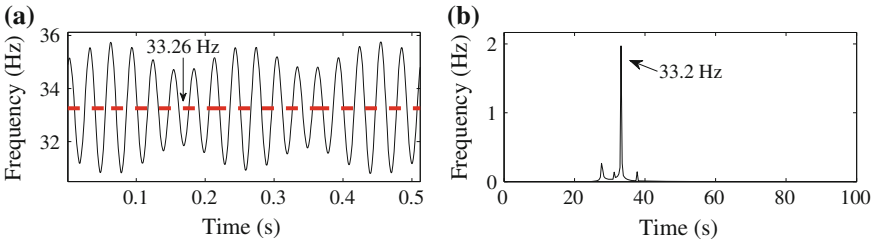
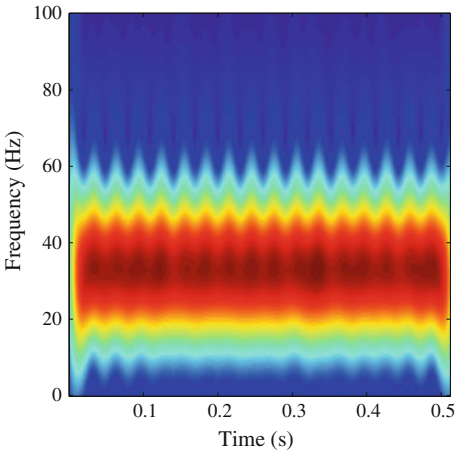


Fig. 21 **a** The extracted IF marked with *blue line* and the mean value marked with the *red line*, **b** the spectrum of the oscillation part

Fig. 22 The TFR of the vibration signal obtained by STFT



rotor system. The spectrum of the oscillation part of this extracted IF is shown in Fig. 21b. The oscillation frequency is also equal to the rotating frequency.

For comparison, the TFR of the vibration signal generated by STFT with $\sigma = 0.01$ is given in Fig. 22 to illustrate the concentration enhancement of the MDT. Moreover, HHT is considered for comparison study. The obtained TFR is presented in Fig. 23. In this case, the HHT spectrum can reveal some oscillation feature.

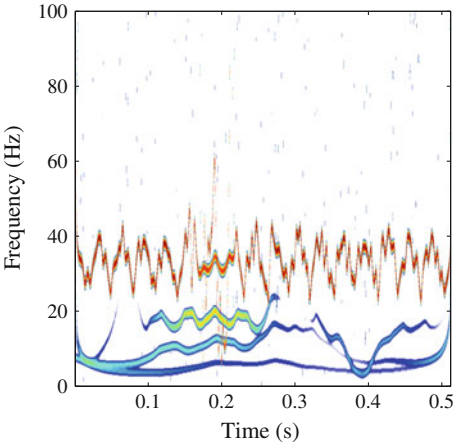


Fig. 23 The TFR of the vibration signal of Bently RK-4 rotor kit obtained by HHT

However, because of the shortcoming of mode mixing, it is incapable of characterizing the true nonlinear IF rule at certain times, such as 0.2 and 0.27 s. Compared with the HHT, the MDT algorithm provides TFR with satisfying energy concentration to represent the nonlinear FM feature and reveal the fault feature of the rub-impact.

7 Applications

In this section, the MDT algorithm will be applied in the feature extraction of vibration signal collected from a heavy oil catalytic cracking machine set with a rub-impact fault. A picture of the machine set and its structure sketch are shown in Fig. 24a, b, respectively. It consists of a gas turbine, compressor, gearbox and a motor. The gas turbine is used to transform heat energy to mechanics energy. Two bearing cases (bushes 1# and 2#) are used to support the gas turbo shaft, another two (bushes 3# and 4#) are used to support the compressor shaft. The hub and the laminas (left component of gas turbo) on the shaft are cantilever. The rotating speed of gas turbo is about 5 800 r/min (96.7 Hz). The instrument of a Bently 3300 system was equipped to monitor its operating condition. Eddy current sensors were mounted on each bearing case in horizontal and vertical directions to capture vibration signals. The sampling frequency is 2 kHz and the sampling number is 1024.

After an overhaul this machine set was running again. It was found that the vibration of bush 2# is over alarm limit, and it is larger than the vibration of bush 1#. The vibration signal of bush 2# and its spectrum are shown in Fig. 25. The operation condition of the machine set became abnormal, and thus it had to be stopped. Except the fundamental harmonic component (96.7 Hz) and its second

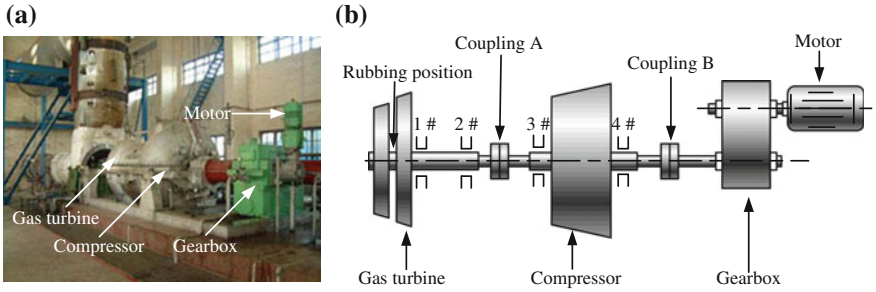


Fig. 24 **a** The picture of the heavy oil catalytic cracking machine set and **b** its structure sketch

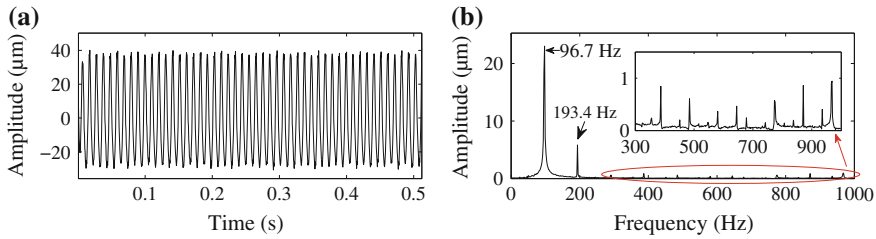


Fig. 25 **a** A vibration signal of the heavy oil catalytic cracking machine set, and **b** its Fourier spectrum

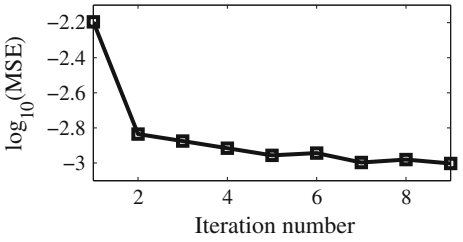


Fig. 26 The logarithmic MSE values of the iterative procedures

harmonic (193.4 Hz) are distinct in the Fig. 25, no other evident fault feature can be found.

MDT is used to analyze the vibration signal and to extract the feature of rub-impact fault. In this case, the value of σ for the initial IF estimation is 0.002, and the threshold is $\delta = 10^{-3}$ for MDT's MSE termination condition. Figure 26 shows the evolution of the logarithmic MSE while implementing iterative procedure of the MDT algorithm. The TFR result by MDT is shown in Fig. 27. It can be clearly found that the FM component has a periodic oscillatory IF. Moreover, the oscillation period is about 10.3 ms, and the corresponding oscillation frequency 96.68 Hz is approximately equal to the rotating speed, as shown in Fig. 28. Consequently, this periodic oscillation feature provides evidence to judge the

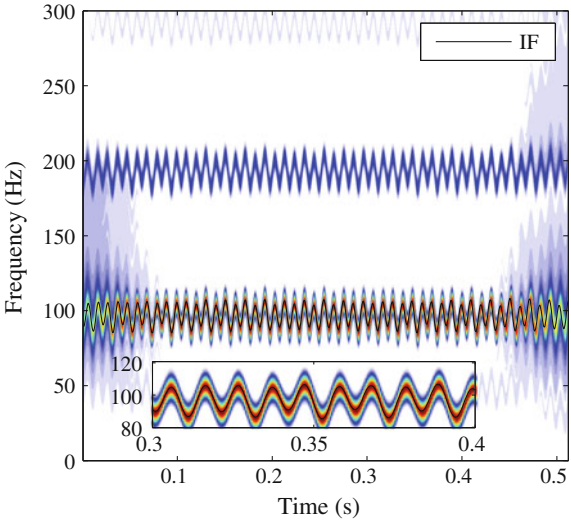


Fig. 27 The TFR of the vibration signal obtained by MDT

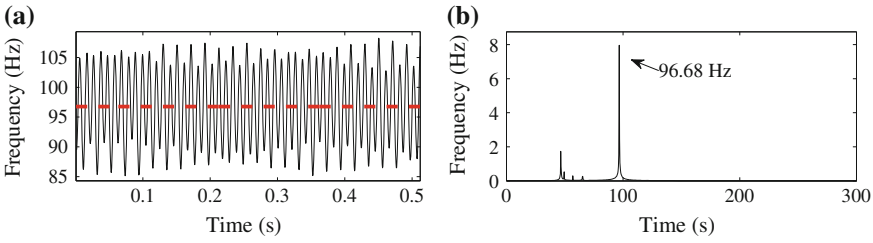


Fig. 28 **a** The extracted IF marked with *blue line* and the mean value marked with the *red line*, and **b** the spectrum of the oscillation part

existence of the rub-impact fault in the rotor system. The analysis results behave better than HHT shown in Fig. 29.

To further verify the effectiveness of the MDT, another vibration signal sampled some time later is analyzed. The waveform and its spectrum are shown in Fig. 30. The rotating frequency and some harmonic components can be clearly seen in this figure. Compared with the vibration signal in Fig. 25, this signal has more complex harmonic components. Moreover, the rotating frequency (25.4 Hz) of the low speed shaft of the gearbox also can be seen in this figure.

The MDT algorithm is used to analyze this measured vibration signal. The TFR obtained by the MDT for multicomponent signal is shown in Fig. 31. The energy in the TF plane is well concentrated around the harmonic components. The FM components could be found clearly in Fig. 31a, where the oscillation phenomena of all IFs of the fundamental component and harmonic components can be observed.

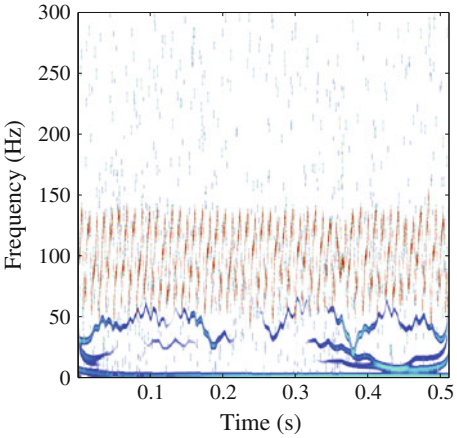


Fig. 29 The comparison study of the vibration signal. The TFRs obained by HHT

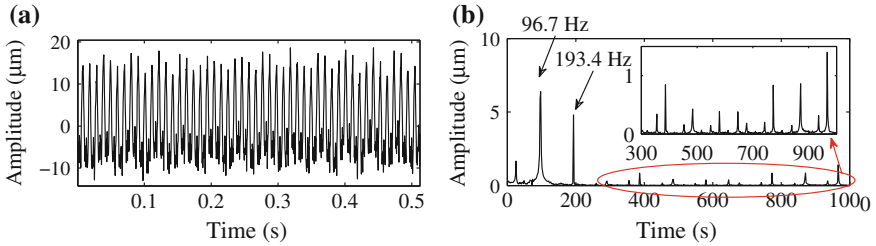


Fig. 30 **a** Another vibration signal of the heavy oil catalytic cracking machine set, and **b** its Fourier spectrum

The oscillation period of the second harmonic is approximately 10.3 ms, which is associated with the frequency of the fundamental harmonic component. More important phenomenon is that the modulation periods of the fourth-harmonic and higher-order harmonics are approximately 32 ms which is approximately equal to three times rotating period of the machine set. It means that the oscillation frequency of the fourth-harmonic component is equal to 1/3 of the rotating frequency. This 1/3 sub-harmonic is another key feature of the rub-impact fault. In order to observe more clearly, the zoomed plot about the fourth-harmonic is given in Fig. 31b. Moreover, Fig. 32a redraws the extracted IF and its mean value which is approximately equal to four times the rotating frequency. The spectrum of the oscillation part of this extracted IF is shown in Fig. 32b.

For the purpose of comparison, the TFR of HHT is shown in Fig. 33. The HHT spectrum is also incapable of characterizing the entire oscillation behaviors of the vibration signal in this case.

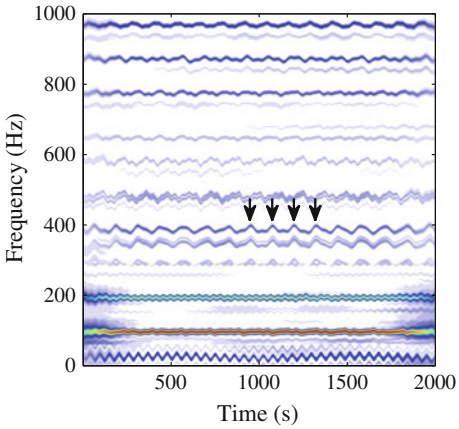


Fig. 31 **a** The TFR obtained by MDT and **b** the zoomed plot about the fourth-harmonic component

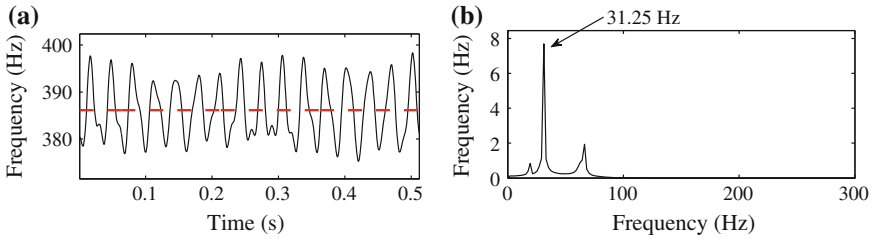


Fig. 32 **a** The extracted IF marked with *blue line* and the mean value marked with the *red line*, **b** the spectrum of the oscillation part

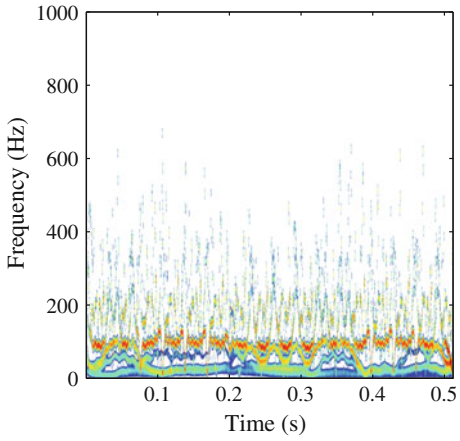


Fig. 33 The comparison study of the vibration signal. The TFR obtained by HHT

In the overhaul afterward, it was found that a rub-impact fault did exist between the hub (rotating with the rotor) and the gas seal (static element). The rubbing position is shown in Fig. 24. The reason is that thermal expansion caused by the increase of the lube temperature elevates the shaft position, and then induces rub-impact with the gas seal. After overhaul, the rubbing area is ground, and the rub-impact becomes slight, and thus the vibration of bush 2# decreased and becomes normal. Thus, the validity of the proposed MDT method is demonstrated.

8 Conclusions

The MDT algorithm provides an iterative TFA method to generate a TF representation with satisfactory energy concentration for both monocomponent signals and multicomponent signals. With the implementation of the iterative procedure, the MDT gradually matches the true IF of the signal, and the energy concentration of TF representation is enhanced and centered around the true IF. Compared with conventional parametric TFA methods, the concentration of the MDT's result is enhanced and the IF estimation accuracy is essentially improved. Moreover, because the MDT is a linear TFA method, it can reconstruct individual components from a multicomponent signal's TF representation.

Then, the validity and practicability of the proposed method are demonstrated by the simulation study and the experiment study, and further by the application in fault feature extraction of a heavy oil catalytic cracking machine set with a rub-impact fault. The simulation signal is used as an example to illustrate the iterative convergence of the MDT algorithm for strong frequency modulation signal. The comparison results indicate that the MDT algorithm behaves better than the other compared methods in providing the TF representation of better energy concentration and achieving more accurate IF estimation for the nonlinear FM signals. The application results show that the MDT method behaves better than HHT and SWT in extracting the feature of highly oscillatory FM signal.

Acknowledgements The authors gratefully acknowledge the National Key Basic Research Program of China under Grant 2015CB057400, the National Natural Science Foundation of China under Grand 51605366 and 51421004, the China Postdoctoral Science Foundation under Grand 2016M590937, and the open fund of State Key Laboratory for Manufacturing Systems Engineering (Xi'an Jiaotong University) under Grand sklms2016004.

References

1. Yan R., Gao R.X., "Energy-based feature extraction for defect diagnosis in rotary machines," IEEE Transactions on Instrumentation and Measurement, 2009, 58 (9): 3130–3139.
2. Randall R.B., Antoni J., "Rolling element bearing diagnostics-A tutorial," Mechanical Systems and Signal Processing, 2011, 25 (2): 485–520.

3. Huang N.E., Shen. Z., Long S.R., et al. “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,” Proceedings of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences, 1998, 454 (1971): 903–995.
4. Yan R., Gao R.X., “Hilbert–Huang transform-based vibration signal analysis for machine health monitoring,” IEEE Transactions on Instrumentation and Measurement, 2006, 55 (6): 2320–2329.
5. Wang Y.X., He Z.J., Zi Y.Y., “A demodulation method based on improved local mean decomposition and its application in rub-impact fault diagnosis,” Measurement Science and Technology, 2009, 20 (2): 025704 (025710 pp).
6. Lei Y., Lin J., He Z., et al. “A review on empirical mode decomposition in fault diagnosis of rotating machinery,” Mechanical Systems and Signal Processing, 2013, 35 (1–2): 108–126.
7. Yan R., Gao R.X., Chen X., “Wavelets for fault diagnosis of rotary machines: A review with applications,” Signal Processing, 2014, 96: 1–15.
8. Wang S., Huang W., Zhu Z.K., “Transient modeling and parameter identification based on wavelet and correlation filtering for rotating machine fault diagnosis,” Mechanical Systems and Signal Processing, 2011, 25 (4): 1299–1320.
9. Chen X., Li X., Wang S., et al., “Composite damage detection based on redundant second-generation wavelet transform and fractal dimension tomography algorithm of Lamb wave,” IEEE Transactions on Instrumentation and Measurement, 2013, 62 (5): 1354–1363.
10. Feng Z., Liang M., Chu F., “Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples,” Mechanical Systems and Signal Processing, 2013, 38 (1): 165–205.
11. Peng Z., Meng G., Chu F., et al., “Polynomial chirplet transform with application to instantaneous frequency estimation,” IEEE Transactions on Instrumentation and Measurement, 2011, 60 (9): 3222–3229.
12. He Q., “Time–frequency manifold for nonlinear feature extraction in machinery fault diagnosis,” Mechanical Systems and Signal Processing, 2013, 35 (1–2): 200–218.
13. Feng Z., Chen X., Liang M., “Iterative generalized synchrosqueezing transform for fault diagnosis of wind turbine planetary gearbox under nonstationary conditions,” Mechanical Systems and Signal Processing, 2015, 52–53 (0): 360–375.
14. Cai G., Chen X., He Z., “Sparsity-enabled signal decomposition using tunable Q-factor wavelet transform for fault feature extraction of gearboxm,” Mechanical Systems and Signal Processing, 2013, 41 (1–2): 34–53.
15. Chen X., Du Z., Li J., et al., “Compressed sensing based on dictionary learning for extracting impulse components,” Signal Processing, 2014, 96: 94–109.
16. Fan W., Cai G., Zhu Z.K., et al., “Sparse representation of transients in wavelet basis and its application in gearbox fault feature extraction,” Mechanical Systems and Signal Processing, 2015, 56: 230–245.
17. He W., Ding Y., Zi Y., et al., “Sparsity-based algorithm for detecting faults in rotating machines,” Mechanical Systems and Signal Processing, 2015.
18. Hu N., Chen M., Wen X., “The application of stochastic resonance theory for early detecting rub-impact fault of rotor system,” Mechanical Systems and Signal Processing, 2003, 17 (4): 883–895.
19. Boashash B., “Estimating and interpreting the instantaneous frequency of a signal–part I. Fundamentals,” Proceedings of The IEEE, 1992, 80 (4): 520–538.
20. Boashash B., “Estimating and interpreting the instantaneous frequency of a signal–part II. Algorithms and applications,” Proceedings of The IEEE, 1992, 80 (4): 540–568.
21. Zhang Y., Obeidat B.A., Amin M.G., “Spatial polarimetric time-frequency distributions for direction-of-arrival estimations,” IEEE Transactions on Signal Processing, 2006, 54 (4): 1327–1340.
22. Liu B., “Instantaneous frequency tracking under model uncertainty via dynamic model averaging and particle filtering,” IEEE Transactions on Wireless Communications, 2011, 10 (6): 1810–1819.

23. Georgakis A., Stergioulas L., Giakas G., "Fatigue analysis of the surface EMG signal in isometric constant force contractions using the averaged instantaneous frequency," IEEE Transactions on Biomedical Engineering, 2003, 50 (2): 262–265.
24. Chen B., Zhang Z., Sun C., et al., "Fault feature extraction of gearbox by using overcomplete rational dilation discrete wavelet transform on signals measured from vibration sensors," Mechanical Systems And Signal Processing, 2012, 33: 275–298.
25. Mallat S. A Wavelet Tour of Signal Processing, San Diego (CA): Academic press, 1998.
26. Qian S., Chen D., "Joint time-frequency analysis," IEEE Signal Processing Magazine, 1999, 16 (2): 52–67.
27. Sejdić E., Djurović I., Jiang J., "Time–frequency feature representation using energy concentration: An overview of recent advances," Digital Signal Processing, 2009, 19 (1): 153–183.
28. Barkat B., Boashash B., "Instantaneous frequency estimation of polynomial FM signals using the peak of the PWVD: Statistical performance in the presence of additive Gaussian noise," IEEE Transactions on Signal Processing, 1999, 47 (9): 2480–2490.
29. Barkat B., "Instantaneous frequency estimation of nonlinear frequency-modulated signals in the presence of multiplicative and additive noise," IEEE Transactions on Signal Processing, 2001, 49 (10): 2214–2222.
30. Li X., Bi G., Stankovic S., et al., "Local polynomial Fourier transform: A review on recent developments and applications," Signal Processing, 2011, 91 (6): 1370–1393.
31. Pei S.C., Huang S.G., "STFT with adaptive window width based on the chirp rate," IEEE Transactions on Signal Processing, 2012, 60 (8): 4065–4080.
32. Olhede S., Walden A., "A generalized demodulation approach to time-frequency projections for multicomponent signals," Proceedings of the Royal Society A, 2005, 461: 2059–2179.
33. Wang S., Chen X., Cai G., et al., "Matching demodulation transform and synchrosqueezing in time-frequency analysis," IEEE Transactions on Signal Processing, 2014, 62 (1): 69–84.
34. Wang S., Chen X., Li G., et al., "Matching demodulation transform with application to feature extraction of rotor rub-impact fault," IEEE Transactions on Instrumentation and Measurement, 2014, 63 (5): 1372–1383.
35. Sejdic E., Djurovic I., Stankovic L., "Quantitative performance analysis of scalogram as instantaneous frequency estimator," IEEE Transactions on Signal Processing, 2008, 56 (8): 3837–3845.
36. Stankovic L., Dakovic M., Ivanovic V., "Performance of spectrogram as IF estimator," Electronics Letters, 2001, 37 (12): 797–799.
37. Ioana C., Zhang Y.D., Amin M.G., et al., "Time-frequency characterization of micro-multipath signals in over-the-horizon radar," IEEE Radar Conference, 2012: 0671–0675.
38. Zhang Y., Amin M.G., Frazer G.J., "High-resolution time-frequency distributions for manoeuvring target detection in over-the-horizon radars," IET Proceedings on Radar, Sonar and Navigation, 2003: 299–304.
39. Yang Y., Peng Z., Meng G., et al., "Characterize highly oscillating frequency modulation using generalized Warblet transform," Mechanical Systems and Signal Processing, 2012, 26: 128–140.
40. Katkovnik V., Stankovic L.J., "Instantaneous frequency estimation using the Wigner distribution with varying and data-driven window length," IEEE Transactions on Signal Processing, 1998, 46 (9): 2315–2325.
41. Stankovic L., "A method for time-frequency analysis," IEEE Transactions on Signal Processing, 1994, 42 (1): 225–229.
42. Orovic I., Stankovic S., Thayaparan T., et al., "Multiwindow S-method for instantaneous frequency estimation and its application in radar signal analysis," IET Signal Processing, 2010, 4 (4): 363–370.
43. Daubechies I., Lu J., Wu H-T., "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool," Applied and Computational Harmonic Analysis, 2011, 30 (2): 243–261.

44. Li C., Liang M., "Time-frequency signal analysis for gearbox fault diagnosis using a generalized synchrosqueezing transform," Mechanical Systems and Signal Processing, 2012, 26: 205–217.
45. Li C., Liang M., "A generalized synchrosqueezing transform for enhancing signal time–frequency representation," Signal Processing, 2012, 92 (9): 2264–2274.

Compressive Sensing: A New Insight to Condition Monitoring of Rotary Machinery

Gang Tang, Huaqing Wang, Yanliang Ke and Ganggang Luo

Abstract With the development of rotary machinery condition monitoring, challenges have often been encountered due to the cumbersome nature of data monitoring. Common methods in signal processing are primarily based on the Shannon sampling principle, which requires substantial amounts of data to achieve the desired accuracy from on-line monitoring signals. This limits their applications in cases for which only small samples can be collected, or cases for which too much data are generating which needs to be largely reduced with under-sampling. Using the Shannon sampling principle, it seems impossible to significantly reduce the quantity of data while preserving adequate useful information for condition monitoring. A newly developed theory termed compressive sensing provides a new insight to condition monitoring and fault diagnosis. It states that a signal can be perfectly recovered from under-sampled data, which means that useful condition information can still be represented by small samples. This study presents novel methods for rotary machinery fault detection from compressed vibration signals inspired by compressive sensing, which can largely reduce the data collection and detect faults of rotary machinery from only a few signal samples. This will greatly help reduce the amount of monitoring data while still guaranteeing a high accuracy of fault detection. Case studies related to roller bearing fault signals are also presented in this study to illustrate the effectiveness of the present strategy.

1 Introduction

As a highly important piece of equipment in various industrial fields, rotary machinery is integral for ensuring security and stable operations of mechanical systems. The rotor and its rotating parts are the two main components of rotary machinery. Critical consequences may result from failures in rotary machinery or its

G. Tang · H. Wang (✉) · Y. Ke · G. Luo
Beijing University of Chemical Technology, Beijing, People's Republic of China
e-mail: hqwang@mail.buct.edu.cn

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors, Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_8

rotating units, and can be more detrimental in the absence of adequate monitoring. In addition, since condition-based maintenance is often required in the modern machinery management, it is necessary to perform condition monitoring to eliminate excess maintenance and guarantee a safe operation. Based on this, researches related to fault diagnosis of rotary machinery have attracted great interests from both academic and industrial communities.

In the past, fault diagnosis was often achieved via equipment disassembly at the job site by experienced maintenance engineers. This manual method presented challenges in guaranteeing high accuracy and efficiency. Presently, with the rapid development of information technology, it is possible to achieve online monitoring and fault diagnosis of rotary machinery. Various unstable factors may exist during the operation of rotary machinery. Abundant status information consistently results from vibrations, allowing vibration signal analysis to be a common and effective method for condition monitoring. Vibration signal analysis is typically performed in the time domain, frequency domain or time-frequency domain [1, 2]. Statistical parameters are adopted to detect and predict faults in the time domain. This method is easily implemented for fault detection, however, it cannot distinguish fault types with high precision [3]. Frequency analysis may be applied to extract fault features [4, 5], identifying fault types by highlighting characteristic fault frequencies in a spectral domain. Signals acquired by sensors, however, often contain noise, thereby complicating effective fault features extraction. Time-frequency methods have been developed to solve these issues, e.g., empirical mode decomposition [6–8] and wavelet analysis [9, 10], and are generally based on the Shannon sampling theory in which the sample frequency must be twice the maximum frequency. This theory indicates that a large amount of data must then be collected, creating an exceptional challenge for signal acquisition, transmission and processing. In addition, since the development of database technology and the automation improvements of large essential equipment, a real-time monitoring techniques have been widely applied. Using this approach, operation data can be acquired by a distributed control system with a high-rate collection, e.g., one data point per microsecond or even higher, to monitor the changes of displacement, acceleration or other parameters. Finally, a large amount of data is collected, and a large scale database or data warehouse is built to improve the accuracy of monitoring and its automaticity.

However, the observed data and parameters are often disorderly and unsystematic, i.e., the features are not obvious for condition monitoring. Meanwhile, complex equipment often generates a large-scale data set to be analysed. Generally, an intelligent fault diagnosis system is an information processing system, which collects a large set of information about an object with the aid of technologies related to sensing, information and data transmission [11]. It must be able to accommodate a lot of original fault information, however, it may also encounter problems of low quality data that could potentially result in uncertain information. Especially, the problems of incomplete information are also exacerbated by the limitations of current data acquisition and monitoring techniques as well as the diverse information of rotary machinery. The incompleteness and discordance of the data presents new challenges to fault diagnosis and condition monitoring. For the fault diagnosis,

incomplete information primarily refers to missing attributes, incomplete data or uncertain information, which would result in an inaccurate conclusion regarding a machine's status. Therefore, how to deal with incomplete monitoring data and make a reasonable inference about a machine's running condition has become a hot topic in intelligent monitoring of rotary machinery.

Moreover, there is a high requirement on the real-time performance for condition monitoring of modern rotary machinery. It is expected that a fault can be discovered once it appears. However, in a big data set generated by continuous monitoring, there is only a small number of data related to a machine's abnormality, since the large majority is healthy and stable information. Thus, it would be much easier if the monitoring data were greatly reduced, while preserving the most useful information. In this case, the pressures on acquisition and post-processing can be relieved with a guarantee to the diagnosis speed and accuracy. Whereas a large amount of sampled data is required to be within the limits of traditional Shannon sampling principle [12] used for perfect post-processing of the observed data. Therefore, it seems impossible to achieve condition monitoring of rotary machinery from an abbreviated data set as suggested above.

Compression of large-scale monitoring data to detect fault features directly from sparse samples is one way to address these challenges. A theory termed compressive sensing is such a way that provides a new insight for solving the above problems, i.e., condition monitoring from compressed samples or incomplete big data sets. The theory states that it is still possible to recover a signal from only a few samples, even with under-sampled incomplete data [13, 14]. It is a big breakthrough in the signal processing field and great attentions have been placed on it since its original proposal. Compressive sensing has been widely applied in various fields, e.g., magnetic resonance imaging [15], seismic wave processing [16] over time, yet many of the studies reported are associated with signal or image reconstruction.

For condition monitoring of rotary machinery, according to compressive sensing, operation information is possible to be reserved with well-designed sampling, then it is possible to store and transmit a small amount of samples and reconstruct them on the receiving side, and detect the fault features from only a few samples. Moreover, the fault features usually can be identified far before signal recovery is complete, thus it is not necessary to recover the signal perfectly. Effectiveness of statistical inference based on compressive sensing has been verified in references [17–20] in related fields, suggesting the possibility of estimating certain characteristic parameters from only a few compressed measurements without ever recovering the actual signals.

In the field of condition monitoring, there are also some related reports found in the relevant literature [21, 22]. Chen et al. [23] built a learning dictionary frame to extract a fault-impact signal. Zhang et al. [24] performed a preliminary study on compressive detection issues of bearing faults. Tang et al. [25] developed a sparse classification method for rotating machinery faults based on a compressive sensing strategy. Results of these studies validate the effectiveness of compressive sensing in machinery fault diagnosis; however, focuses were primarily on sparse representation or reconstruction of fault signals.

Considering the complexity of both condition monitoring and compressive sensing, there are still many obstacles that must be overcome, especially on the extraction of fault features from compressed signals. The motivation of this paper is to briefly introduce the compressive sensing theory and present some applications for the condition monitoring of rotary machinery. These compressive-sampling-based methods can help to promote the fault detection efficiency of rotary machinery faults from under-sampled signals, which will provide new insights to this research fields.

In this paper, roller bearings are used as an example to explain the main concept of the proposed strategy [26, 27]. Statistical inference based on compressive sensing has been studied in other fields [28–30] as mentioned above, yet there are still many obstacles to be overcome when applied to bearing fault detection. The bearing fault signal consists of impulses and in the commonly utilized Fourier or wavelet domain, its sparsity does not completely meet the requirements of compressive sensing, thereby increasing difficulty of the compressive sensing process. Also yet to be resolved also is the identification of bearing fault features to be extracted from under-sampled signals and the integration process for compressive sensing into the bearing fault diagnosis. In this study, we try to develop applicable condition monitoring strategies for bearing faults from under-sampled vibration signals, and perform simultaneously sampling and detection without a complete recovery of the incomplete signal.

The rest of this paper is organized as follows. Section 2 states the fault detection problems in rotary machinery monitoring. Section 3 provides a brief introduction to compressive sensing. Section 4 shows three proposed methods and case studies for bearing fault detection with simulation and experiments. Conclusion is drawn in Sect. 5.

2 Problem Statement

In the condition monitoring of rotary machinery, to reveal the operation status accurately and comprehensively, a large number of signals are often collected, including the signal of operating condition (e.g., speed, pressure), the vibration signals, the surrounding signals (e.g., temperature), etc. This leads to mutual crosslinking, which complicates the relationship between different signals. The intensity trends of a vibration signal are often related to the operation state of a piece of equipment, thus they are important indicators of whether a machine is running properly or not. Furthermore, the intensity is also closely related to the working condition and the surroundings. They are closely linked to each other, therefore none are dispensable. However, the limitations of the field environment, often result in a lack of data, which adversely affects the judgment of the machine status. In addition, the complication of a piece of equipment usually causes a complex signal transmission path, which leads to a serious noise interference, or

even incorrect signals. Thus we have to pre-processing the observed signals, e.g., eliminating invalid signals, which often renders the data incomplete.

In short, the big data related to rotary machinery are often interfered by the surroundings, which makes the data difficult comprehensively acquired. Therefore, it is necessary to develop a strategy to deal with the big but incomplete monitoring data. Ideally, the big data should be compressed or compressively sensed without losing important information.

Without a loss of generality, a simple detection issue of bearing faults can be formulated as

$$x = s + n \tag{1}$$

where s is a known signal of interest, x denotes the observation signal, and n is mixed noise with interference signals from surrounding devices.

Provided s denotes a vibration signal related to a bearing fault, the fault detection problem then is to distinguish s from x. One of the common methods to distinguish the fault component s from the mixture signal x is to proceed in a transforming domain:

$$y = \Phi x = \Phi\varphi^H\varphi(s+n) = A\varphi(s+n) \tag{2}$$

where x is a signal of $N \times 1$ dimension, Φ is an M × N measurement matrix, $M \leq N$, and each row of Φ represents a sensor to measure x. φ is a $N \times N$ column orthonormal basis matrix, and the superscript φ denotes a conjugate transposition. $A = \varphi^H\varphi$ is often designated the sensing matrix to measure the transformed data $u = \varphi x$. y is a $M \times 1$ measurement vector denoting the observation of $y = Au$. When all N measurements are available, i.e., $M = N$, then, $\Phi^H\Phi = \Phi\Phi^H = I_{N\times N}$, indicating that y is an observation of x with full sampling, which can be solved by many methods.

However, to facilitate data acquisition and bypass the limitations resulting from incomplete and imprecise knowledge, $M \ll N$ is often encountered or expected. y is then indicated as a compressive sensing of signal x. It would be promising if required information of the original signal x could be deduced from the compressed observation y without reconstruction, i.e., the compressed detection problem.

3 Compressive Sensing Theory

3.1 Shannon's Sampling Theory

Shannon's sampling theory was first proposed by Shannon in 1949 [31]. According to the theory, if a continuous signal can be completely represented by a cluster of samples processed at discrete time, then the samples must occur at more than twice the sampling frequency of the highest frequency of the signal.

If the signal xt is sampled through the sampling frequency f_s (sampling interval $T_s = 1/f_s$), then the sequence can be generated at $\{\ldots, x(-nT_s), \ldots, x(-T_s), x(0), x(T_s), \ldots, x(nT_s), \ldots\}$,

$$x_s(t) = \sum_{n=-\infty}^{\infty} x(nT_s) \cdot \delta(t - nT_s) = x(t) \cdot \sum_{k=-\infty}^{\infty} \delta(t - nT_s) \tag{3}$$

where $\delta(t - nT_s) = 1$ at $t = nT_s$, and $\delta(t - nT_s) = 0$ elsewhere. f_s is called the Nyquist frequency [31]. Based on this theory, the maximum frequency in the signal $x(t)$ is $f_s/2$.

The Nyquist frequency must be reached in signal band-limited processing. However, with the development of information technology, the bandwidth of the signals has been so widely expanded that many new troubles arise when dealing with these signals in data collection, data transmission and data storage.

In addition, a lot of unimportant and redundant information is contained in the sampled data. It costs large amounts of time to store and transmit data, in addition to perform an increasing time in signal processing.

3.2 Compressive Sensing

The theory of compressive sensing has been developed in the field of signal processing. It brings a new inspiration to solve problems of big data compression, incomplete data processing and rapid detection from small samples, which is regarded as a breakthrough of the Shannon sampling theorem. Here we give a brief introduction about the theory. For more detail, please refer to [13, 14].

Provided that a perceptual measurement matrix $A = \Phi\varphi^H$ satisfies the isometric constraint conditions, $u = \varphi x$ defines a representation of a sparse signal x as

$$y = Au \tag{4}$$

where x is a $N \times 1$ vector signal, Φ is a $M \times N$ measurement matrix, $M \leq N$, and each row of Φ represents a sensor to measure x. φ is a $N \times N$ column orthonormal basis matrix and the superscript H denotes a conjugate transposition. $A = \Phi\varphi^H$ is often termed the sensing matrix to measure the transformed data $u = \varphi x$, y is a $M \times 1$ measurement vector denoting a compressive sensing of the original full data.

Because $M \leq N$, thus Eq. (4) is an under-determined problem, whose solution can be approximately pursued as

$$\min\|\theta\|_0 \quad \text{s.t.} \quad y = Au = \Phi\varphi^H(\varphi x) \tag{5}$$

Owing to the sparsity promotion strategy, if x is sparse in φ, u and x can be recovered from the small observations y.

The theory employs a sparse space φ to represent the signal x and obtain a small amount of observation data y. In this way, the signal sampling is converted into an information sampling. Then by solving an optimization problem, the original signal x can be recovered from compressed observed data y. With this theory, the sample data no longer depends on the bandwidth of the signal, but on the information structures and contents of the signal. Compressive sensing makes it possible to solve inference problems with low sampling rates.

It also provides a new insight to condition monitoring of rotary machinery. According to the compressive sensing theory, a signal can be represented by or sufficiently approximated to a linear combination of predefined atoms. Then the compression efficiency can be greatly improved and processing costs can be largely reduced. Furthermore, for the condition monitoring of rotary machinery, fault features extraction from small samples are as important as those from continuously measured large samples. If we can detect the faults from only a few under-sampled signals, i.e., overcome the limitations of the traditional Shannon sampling theorem, then the requirements surrounding the data acquisition and post-processing can be greatly reduced, in addition to the time costs of condition monitoring.

Generally, sparse representation, sampling schemes and solutions of underdetermined equations are three key issues for compressive sensing technique.

3.3 Sparse Representation of a Signal

According to the compressive sensing theory, sparse representation of a signal is a precondition to recover the original signal. In many methods for signal compression, the signal is often transformed into another domain first with orthogonal projections. Then only samples at positions with large absolute values in the transform domain are reserved to obtain a compressed signal. This is called sparsity which means that a signal can be represented by a liner combination of a small amount of elements. This signal representation theory was originally developed by Mallat and Zhang with a complete dictionary sparse decomposition [32].

In general, a set of functions $\{\varphi_i\}$ can be found in Hilbert space $L_2(R)$ so that signal y can be expressed as a liner combination of N basis $\{\varphi_i\}$. So,

$$\mathrm{y} = \Phi \mathrm{x} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i} \varphi_\mathrm{i} \tag{6}$$

where x_i is the coefficients of y in dictionary $\Phi = (\varphi_1, \ldots \varphi_N)$. x and y are equivalent representations of the same signal. The difference is that y is in the time domain and x is in the dictionary Φ. We say that signal y is K-sparse, which means that the x_i coefficients in formula (6) has K-nonzero elements. In practice, x is considered to be compressible if there are few large coefficients and many small coefficients.

Besides sparsity, the other key point of sparse representations is incoherence which means that the basis must be obviously different [33]. The coherence between two orthonormal matrixes ϕ_i and ϕ_j is defined as,

$$\mu(\phi_i, \phi_j) = \sqrt{n}\max_{1 \le k,j \le n} \left| \langle \phi_i, \phi_j \rangle \right| \tag{7}$$

There are two main issues in sparse representation: how to build a redundant dictionary and how to design a decomposition method. At present, the dictionaries mainly include local cosine dictionary, over-complete wavelets, curvelets and Gabor dictionary [32]. Furthermore, the decomposition methods mainly include Matching Pursuit (MP) [34], Orthogonal Matching Pursuit (OMP) [35], Basis pursuit (BP) [36] and FOCUSS [37].

3.4 Sampling Method

Sampling is the first step of conducting condition monitoring of rotating machinery. Traditional sampling methods must obey the Nyquist sampling theorem to maintain essential features of the signal so that the analysis results via Fourier transform, wavelet transform and Hilbert transform make sense. This usually generates vast amounts of monitoring data. The sampling frequency based on the compressive sensing theory may be much lower while still maintaining the signal features well.

In compressive sensing theory, the measurement matrix is the key to sampling. To ensure different sparse signals is not projected to the same M-dimensional measurement matrix for the perfect reconstruction of sampled signal, the measurement matrix must satisfy the principle of restricted isometry property (*RIP*) that the measurement matrix Φ is noncoherent with sparse representation basis φ.

RIP can be described below. There is an isometric constant constraint $\varepsilon \in (0, 1)$, that allows the following formula to be true for any K-sparse signal:

$$(1 - \varepsilon)\|\boldsymbol{\Phi\Psi} x\|_2^2 \le (1 + \varepsilon)\|x\|_2^2 \tag{8}$$

Determination of the sampling method is essential to designing a measurement matrix Φ that meets *RIP*. Gaussian random measurement matrix and Bernoulli random measurement matrix are often employed in compressive sensing theory. The former obeys the $N(0, 1)$ normal distribution, and the latter meets the Bernoulli distribution. It has been proved that the Gaussian random matrix can meet *RIP* with great probability. Such an irregular sampling (random sampling) method is simple to design, and usually performs perfectly in the reconstruction of under-sampled data. Therefore, a Gaussian random matrix is employed to conduct compression measurement in condition monitoring and fault diagnosis of rotating machinery.

3.5 Optimization Solving Strategy

The process of compressive measurement can be described as:

$$y = \Phi x = \Phi \varphi^H \varphi(s+n) = Au \tag{9}$$

The sparse solution can be approximately pursued as:

$$\min \|u\|_0 \quad \text{s.t.} \quad y = Au = \Phi \varphi^H (\varphi x) \tag{10}$$

However, $M \leq N$, thus Eq. (10) based on minimum l_0-norm is an under-determined problem within an uncertainty of solutions. Convex-optimum algorithm and greedy algorithm are most common used methods to solve the above issue. Optimization objective is replaced by l_1-norm that can transform the problem into linear programming and insure the uniqueness of the solution, which described as:

$$\min \|u\|_1 \quad \text{s.t.} \quad y = Au = \Phi \varphi^H (\varphi x) \tag{11}$$

Basis Pursuit (BP) is a typical algorithm based on l_1-norm optimization. A local optimal solution is selected to approximate the original signal in each iteration of Greedy algorithms, which has a lower computational complexity than convex-optimum algorithm. Matching Pursuit (MP), Orthogonal Matching Pursuit (OMP) and their improved algorithm are typically employed in the optimization solving.

4 Proposed Strategies and Applications

4.1 Experiments

Experiments are carried out to validate the effectiveness of the proposed method. The test rig and the faulty roller bearings are shown in Fig. 1, which is composed of a motor, a coupling, a rotor and a shaft with two roller bearings. Here we do the experiments with roller bearings with single fault in the outer race, inner race and rolling element, respectively. The fault sizes are all width of 0.7 mm and depth of 0.25 mm. Sample frequency is 100 kHz at a shaft speed of 500, 900 and 1300 rpm, respectively. Vibration sensors are located at positions near bearings to mitigate the effects of signal attenuation. The bearing housing is considered to be a superior location for bearing arrangement. Vibration signals are measured by an accelerometer located at the top of the bearing house and the theoretical values of the fault characteristic frequency are shown in Table 1. All data using in this paper are processed through the normalization.

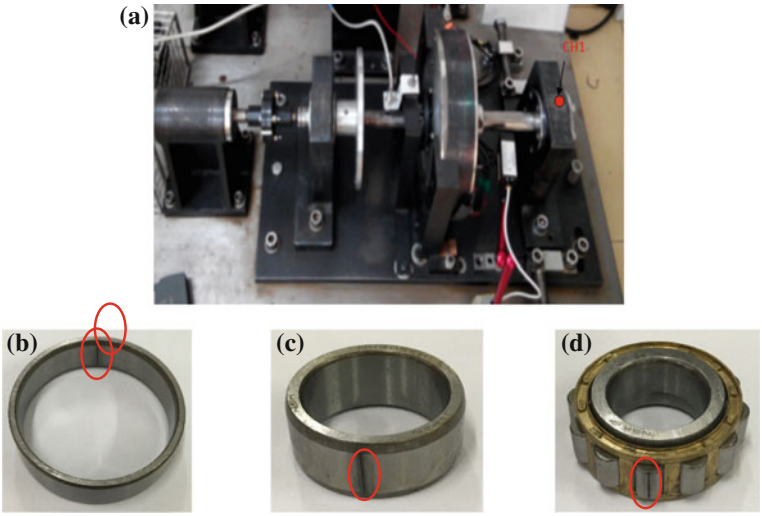


Fig. 1 **a** Fault test rig of roller bearing **b** outer-race fault **c** inner-race fault **d** rolling-element fault

Table 1 Theoretical values of the fault characteristic frequency

Running speed (rpm)	Fault type		
	Outer-race fault (Hz)	Inner-race fault (Hz)	Rolling-element fault (Hz)
500	33.20	56.09	39.33
900	59.76	100.97	70.8
1300	86.32	145.84	102.26

4.2 Reconstruction of Incomplete Vibration Signal

Continuous condition monitoring always leads to big data, which is a major challenge for fault diagnosis. Inspired by the compressive sensing theory, reconstruction from a limited samples provides a new idea for signal storage and transmission. If the original vibration signals can be reconstructed from few samples, it enables the storage of a small amount of samples instead of the whole data set, in addition to the reconstruction of the limited samples to obtain the raw vibration signals when necessary. One of the key preconditions for the compressive sensing theory is that the analyzed signal must be sparse or compressible. Unfortunately, the vibration signals of rotary machinery are often insufficiently sparse in the common transform domain, which presents an obstacle to the application of compressive sensing in fault diagnosis. Here a compression and reconstruction strategy based on compressive sensing is presented to show the potential applications.

Vibration signals measured from faulty bearings are always drowned out by noise, which weakens the sparsity of the vibration signals. Thus, in this section a sparsity-promoted approach based on segmentation threshold denoising is developed.

As shown in Fig. 2, the original vibration signal is first divided into several segments based on its peaks, which are the significant features in faulty vibration signals. Then a threshold is set for denoising, through which the vibration signal becomes sparser. Since the vibration signal becomes adequately sparse, the unit matrix is selected as a sparse matrix, while the Gaussian random matrix is selected as the measurement matrix to gain random observations in order to meet the requirements of compressive sensing. Finally, the signal denoising and recovery are obtained via implementation of a matching pursuit strategy.

A vibration signal of a roller bearing with an outer-race fault operated at 1300 rpm is shown in Fig. 3, which shows that the signal is not significantly sparse. Thus, a sparsity-promoted method based on the segmentation threshold denoising is used to increase the sparsity of the original signals as shown in Fig. 4. After segmentation threshold denoising, the vibration signal becomes much sparser and

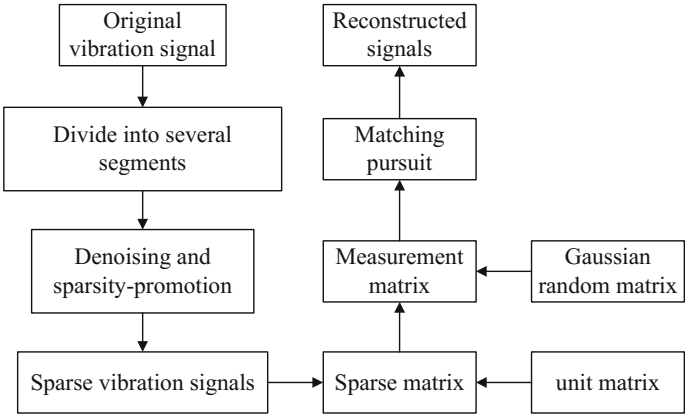


Fig. 2 The flowchart of the proposed compression and reconstruction strategy

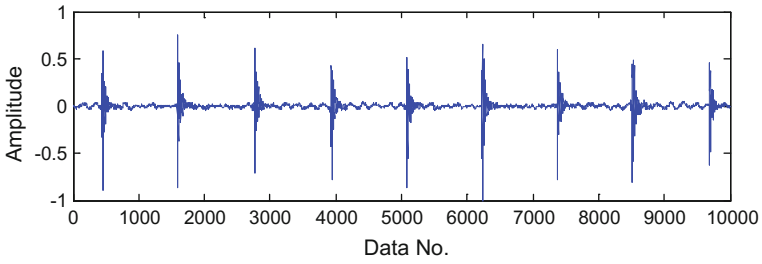


Fig. 3 Original vibration signal of a roller bearing with an outer-race fault operating at 1300 rpm

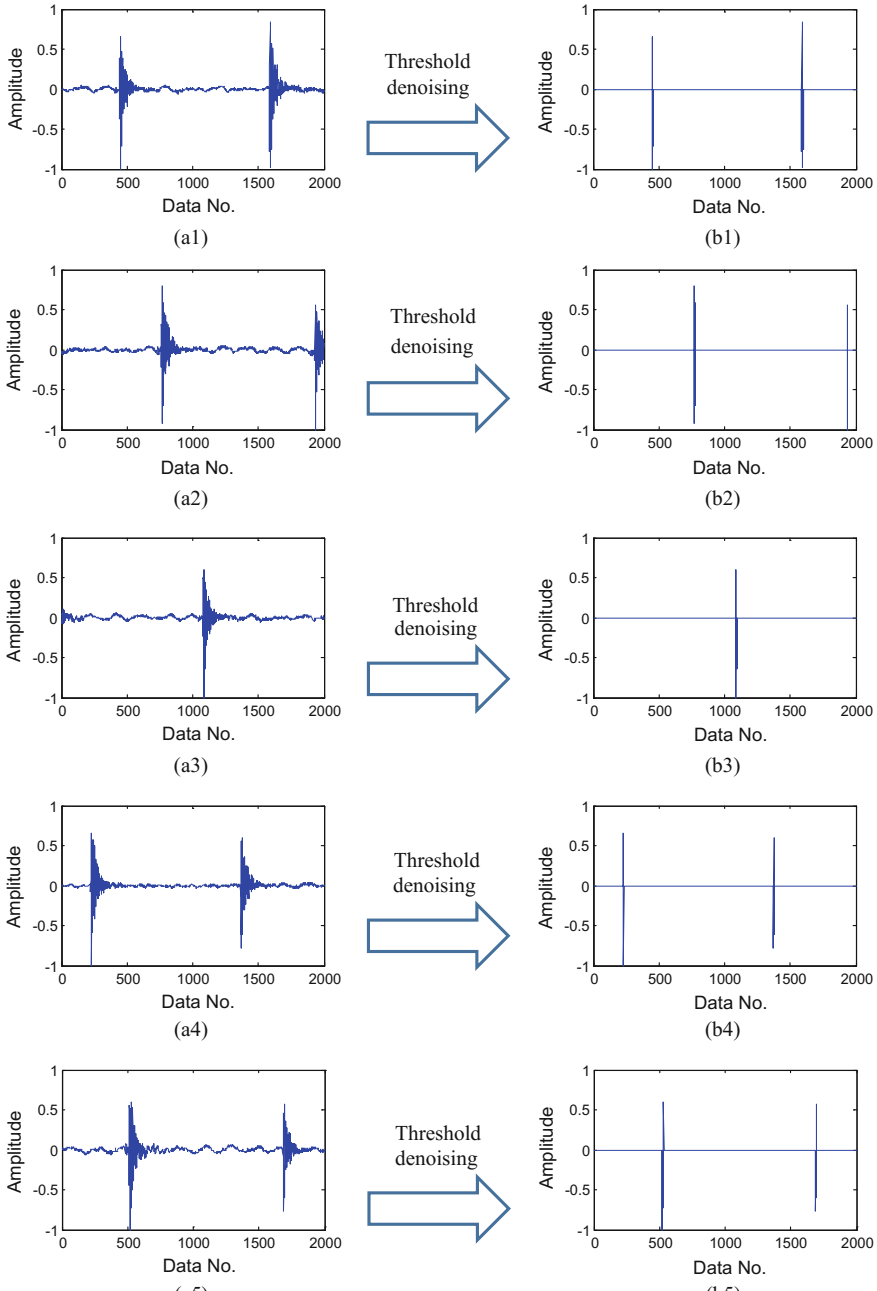


Fig. 4 Segmentation threshold denoising: **a1**–**a5** original segmentation, **b1**–**b5** after segmentation threshold denoising

smoother as shown in Fig. 5. If the signal is sparse enough, the compressive sensing theory can be applied to reconstruct the vibration signal. The dimension-reduced signal with 1000 samples as presented in Fig. 6, is achieved through random sampling. Through the application of a matching pursuit algorithm, the original signal can be recovered as shown in Fig. 7, and the envelope spectrum is shown in Fig. 8, through which the running status of roller bearing can be judged according to the fault characteristic frequency.

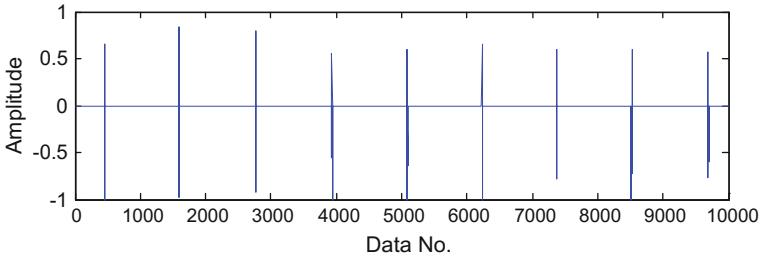


Fig. 5 Vibration signal after segmentation threshold denoising

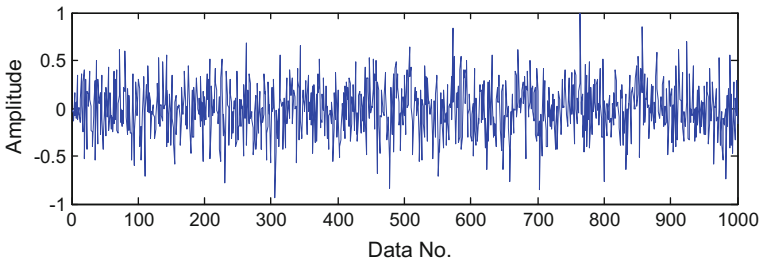


Fig. 6 Compressed sampling with 1000 random samples

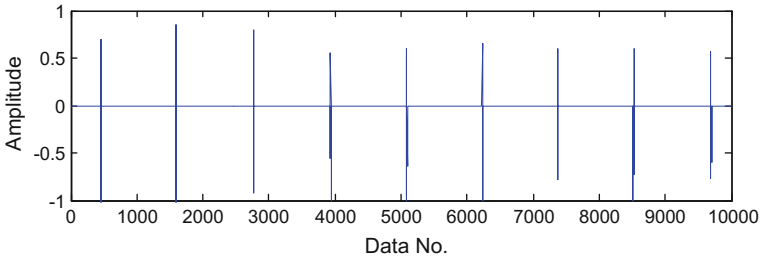


Fig. 7 Reconstructed signal by compressive sensing strategy

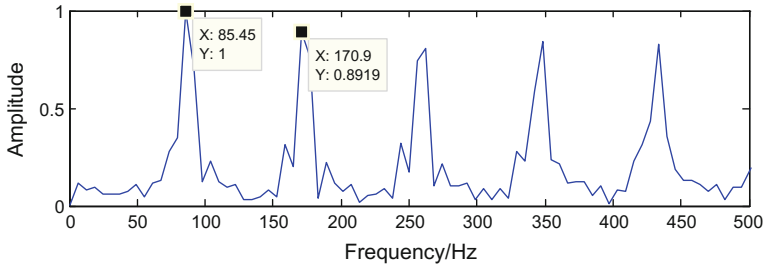


Fig. 8 Envelope spectrum of the recovered signal

4.3 Fault Classification of Rotating Machinery [25]

The compressive sensing theory has proved its capability in reconstruction, de-noising and feature extraction of rotating machinery vibration signal. Additionally, it can be applied to fault diagnosis and classification in the case of partial reconstruction. Bearing signals are taken for example in this section, introducing a rotating machine fault classification method based on dimension reduction sampling and sparse representation.

A redundant dictionary should contain all possible types of signals that any test signal can be described as the linear combination of the vectors in redundant dictionary. For bearing signals, the redundant dictionary consists of the normal signal, inner ring fault signal, outer ring fault signal and rolling elements fault signal, a total of four signal types. E is defined as the redundant dictionary composed of k categories of samples with the following configuration:

$$\begin{aligned} E &= [E_1, E_2, \cdots, E_i, \ldots, E_k] \\ &= [v_{11}, \ldots v_{1N_1}, v_{21}, \ldots, v_{2N_2}, \ldots, v_{i1}, \ldots v_{iN_i}, v_{k1} \ldots, v_{kN_k}] \in R^{M \times N} \end{aligned} \tag{12}$$

$$N = N_1 + N_2 + N_3 + \cdots + N_k \tag{13}$$

where $E_i = [v_{i1}, v_{i1}, \ldots, v_{iN_i}] \in R^{M \times N_i}$ indicates the number of samples N_i of the ith category fault.

After configuration of the redundant dictionary, the test signal sample x of ith category can be described as the linear combination of the vectors in the redundant dictionary,

$$x = Eu \in R^M \tag{14}$$

Thus, the bearing signal x is represented as the sparse vector $u = [0, \ldots, 0, u_{i1}, u_{i2}, \ldots, u_{in_i}, 0, \ldots, 0,]^T \in R^N$ in the transform base E which consists of over-complete training samples.

Gaussian random measurement matrix $R \in R^{D\times M}(D \ll M)$ that contains i.i.d. $\mathcal{N}(0,1)$ entries processes the bearing signal sample x and redundant dictionary E by random mapping dimension reduction, to provide compressive observations $\tilde{y} = Rx$ and sensing matrix $\widetilde{E} = RE$.

$$y = Rx = REu = \widetilde{E}u \in R^D \tag{15}$$

For each test sample x, its sparse solution α of the training set E can be obtained through the SP algorithm. A new set of sparse vector is defined and $u = \sum_i \delta_i(u)$, setting zero value for all elements except those ones corresponding to the ith signal category. Thus, the mapping feature of the test sample x in the ith category has the following formula:

$$\hat{y}_i = \widetilde{E}\delta_i(u) \tag{16}$$

The residual error between compressive observations y and feature value $\hat{y}_i$ is calculated:

$$\min_i r_i(y) = \left\|y - \widetilde{E}\delta_i(u)\right\|_2 \tag{17}$$

The category of the test sample can be determined by the minimum residual error.

The flow chart of this sparse representation classification framework based on compressive sensing is shown in Fig. 9.

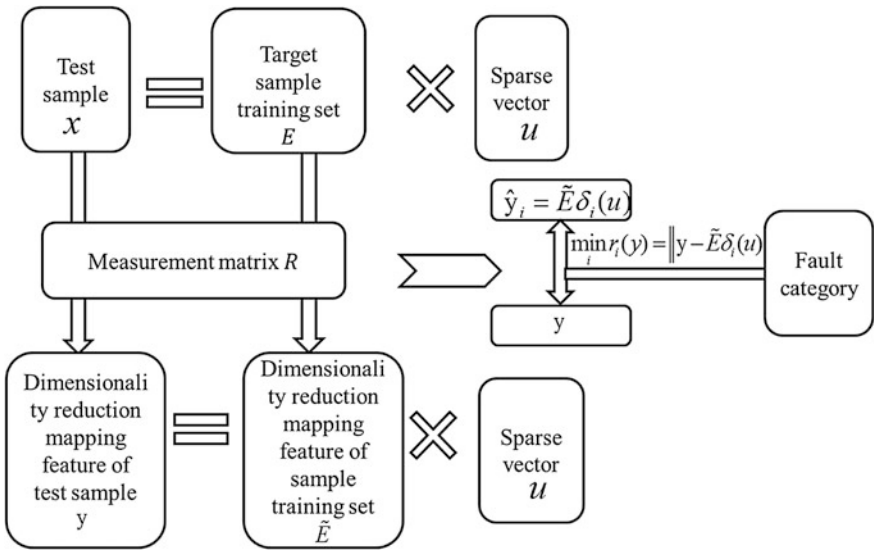


Fig. 9 Flowchart of the sparse representation classification framework

Table 2 Fault classification results at different rotating speeds

Category	Rotating speed (r/min)		
	500	900	1300
Inner race fault	0.996	0.964	0.984
Outer race fault	0.948	1	1
Roller element fault	0.880	0.998	0.976
Average classification accuracy rate	0.941	0.987	0.987

Each fault signal test group is 500 at three different speeds, concluding 500, 900 and 1300 rpm. Fault identification and classification accuracy of the proposed method is presented in Table 2.

Traditional fault pattern recognition methods, such as BP neural network and SVM methods, are usually based on the characteristic parameters of time and frequency domains to achieve fault classification. In this section, a compressed sensing sparse representation classification algorithm (SRC) is proposed. The solution is a sparse vector by SP algorithm and residual error calculation of the feature and observed values, determining the category of the test signal. Comparison analysis is shown in Fig. 10. The SRC method demonstrates its advantage of having a higher accuracy rate in rotating machinery fault classification than traditional BP and SVM pattern recognition methods.

To investigate the effect of the length of original signals on the sparse classification results, the average classification accuracy rate of the proposed SRC algorithm is observed to be a gradually increased trend at different rotating speeds when the length of each input signal is varied from 5000 to 50,000. When the signal

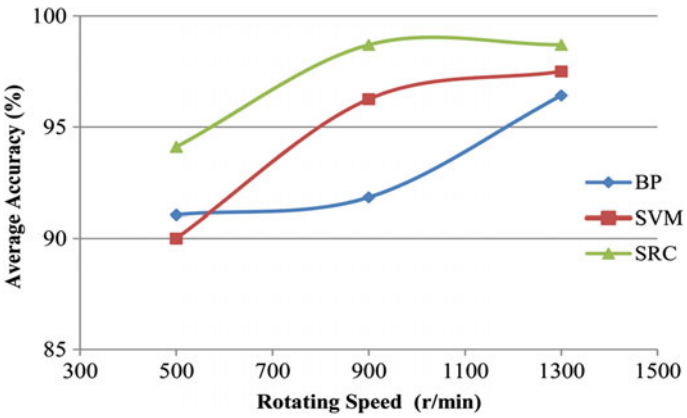


Fig. 10 Classification results of SRC method in comparison with BP and SVM

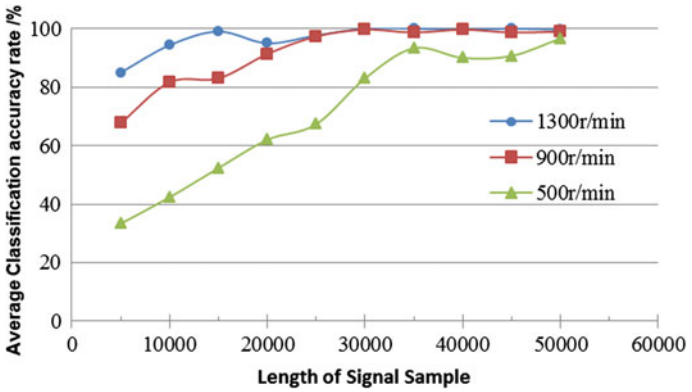


Fig. 11 Effect of the length of signal M on average classification accuracy

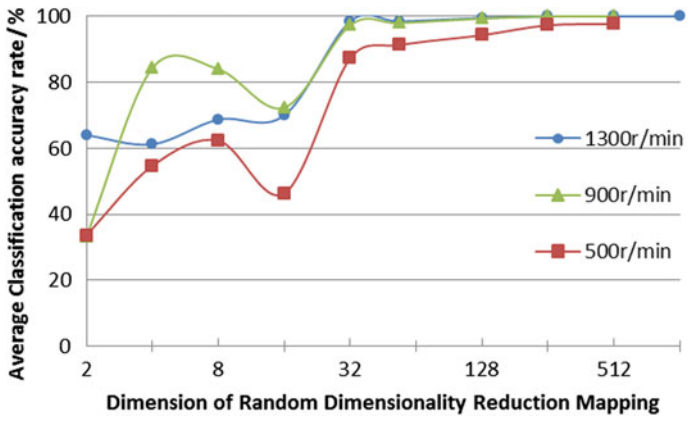


Fig. 12 Influence of variable measurement dimension D on average classification accuracy

length is more than 30,000, the average recognition rate by SRC can reach 99.67%, and more details are shown in Fig. 11.

Sparse dimension reduction parameter D is directly related to the feature extraction and preserving of original signals, thus it is necessary to discuss the influence of this measurement dimension D on classification accuracy rate of the proposed SRC. Figure 12 indicates that the average classification accuracy increases as the $D = 2^j (j = 1, 2, \ldots, 8)$, and when $D \geq 2^5 = 32$, the classification accuracy of the three kinds of bearing faults can reach 98.5%.

4.4 Compressive Sensing of Bearing Fault via Characteristic Harmonic Detection

As mentioned above, the vibration signal of a roller bearing is insufficiently sparse in the Fourier domain to meet the requirement of compressive sensing. It is well known that when a defect occurs in roller bearing, an impulse will be generated when the bearing strikes another surface and periodic impulses will be generated, termed fault characteristic frequency. The inadequate sparsity of a vibration signal exerts a negative effect on a perfect signal reconstruction. If the gathered data is incomplete or compressed due to some reasons, it should be recovered first before using post-processing methods to identify the condition of roller bearing. This makes it difficult to meet the efficiency requirement of real-time condition monitoring. Ideally, the fault detection would be performed with compressed samples directly without complete recovery [26].

The sparsity of a signal is regarded as a priori information for most existing reconstruction algorithms based on compressive sensing theory. However, in practice, it is difficult to achieve a perfect sparse representation and obtain a specific sparsity. Therefore, the sparsity of a signal must be estimated correctly, otherwise, the sparsity of a signal will be an obstacle to the application of compressive sensing.

In fault diagnosis of roller bearing, the objective is to extract fault features rather than data reconstruction. Thus, complete reconstruction of a signal is not necessary in all cases. To our knowledge, the envelope signal of a roller bearing consists of a variety of harmonic waves as sub-components, which are related to the fault features. In addition, it is well known that the sparsity of a harmonic wave in the Fourier domain has a value of 2. If we can detect these harmonic waves related to the fault features in Fourier domain, a decision as to whether or not a fault exists in the roller bearing can be made [26]. Based on this idea, a compressed fault detection method for roller bearing is developed in this work and the fault detection flowchart is presented in Fig. 13.

Here the Fourier basis is selected for sparse representation, and a Gaussian random matrix is chosen as a measurement matrix to reduce the amount of bearing vibration signal. Finally, the matching pursue algorithm, such as orthogonal matching pursue (OMP), compressive sensing sampling matching pursue (CoSaMP), is utilized to detect the harmonic wave with frequencies of interest.

The proposed detection strategy is implemented to extract the fault features with a fault on the inner race at a shaft speed of 900 rpm. The waveform with impulses in time domain is presented in Fig. 14. In generally, it is difficult to extract fault characteristic frequencies from such a large number of samples. Therefore, the proposed compressed fault detection method is applied to extract the fault features. As mentioned above, the Gaussian random matrix is selected as a measurement matrix while Fourier basis is chosen for sparse representation. Next, the detection method based on CoSaMP is used to extract the fault characteristic frequency,

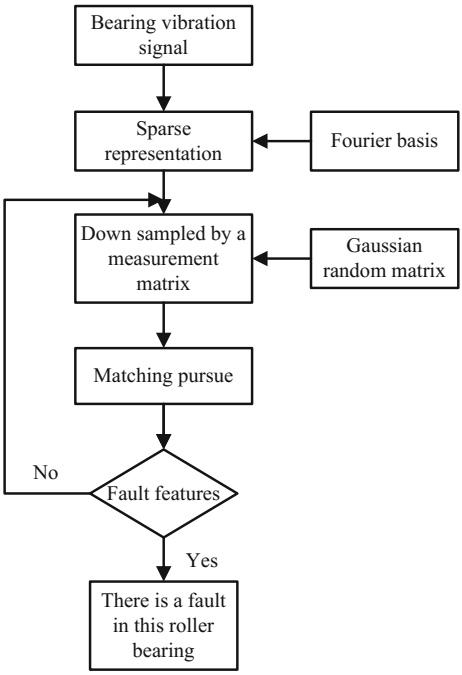


Fig. 13 Scheme of the proposed fault detection strategy with compressive sensing of characteristic harmonic waves

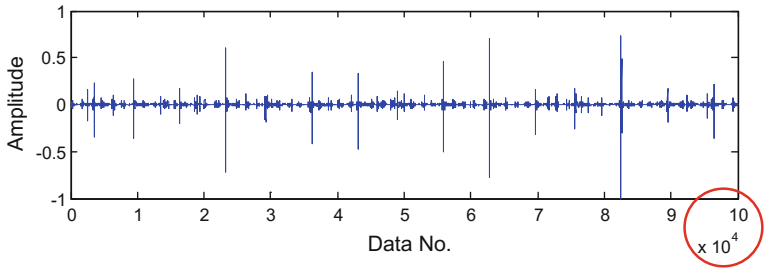


Fig. 14 Time domain waveform of a roller bearing with a fault on the inner-race at 900 rpm

where the sparsity K is set to 2. With a measurement matrix, the number of samples could be compressed to 800 as shown in Fig. 15. The frequency of the first detected harmonic component is 100.6 Hz, as shown in Fig. 16, which is almost equal to the theoretical value. Furthermore, the value twice to the fault characteristic frequency can also be determined, as shown in Fig. 17. Therefore, it could be concluded that a fault existed on the inner race. Different dimension of 400 is utilized to fully validated the effective of the proposed method. From the results in Figs. 18, 19 and 20, a conclusion can be drawn that the method proposed in this work can also detect the faults with 400 observations.

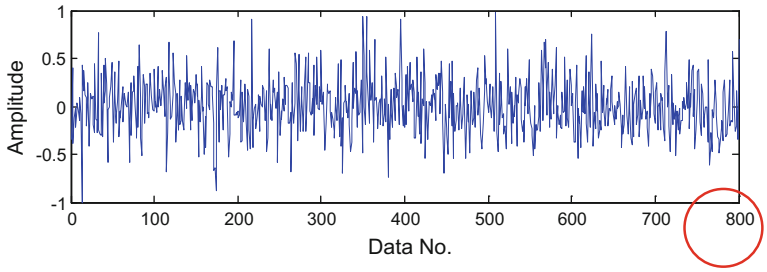


Fig. 15 Random sampling through compressed sensing with 800 observations

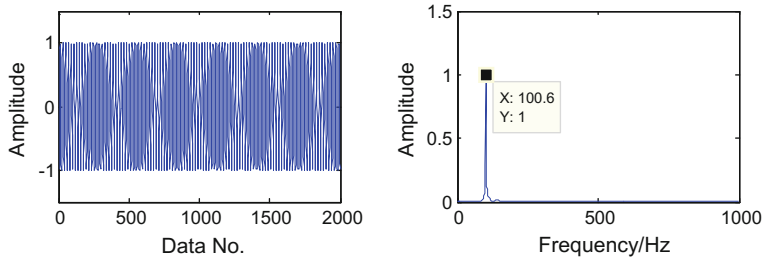


Fig. 16 Fault characteristic frequency of the first detected harmonic component from 800 samples

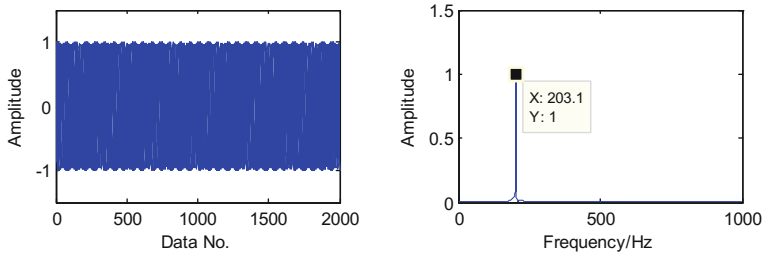


Fig. 17 2 * Fault characteristic frequency of the detected harmonic wave from 800 samples

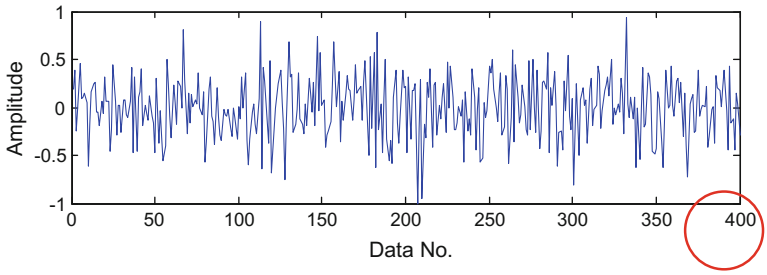


Fig. 18 Random sampling through compressed sensing with 400 observations

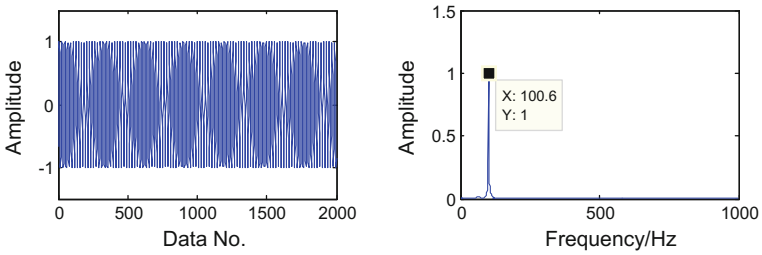


Fig. 19 Fault characteristic frequency of the first detected harmonic component from 400 samples

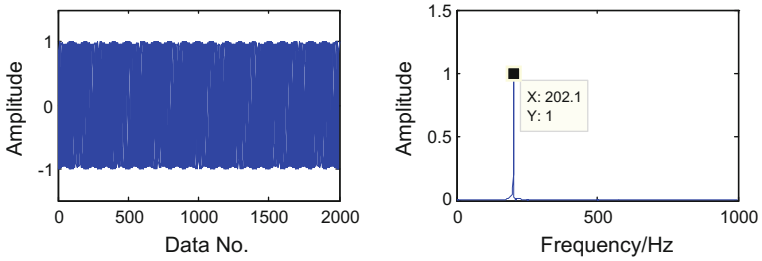


Fig. 20 2 * Fault characteristic frequency of the detected harmonic wave from 400 samples

5 Conclusions

To solve the problems of big data and incomplete small samples in condition monitoring of rotary machinery, this paper introduced a newly developed compressive sensing theory to the field of rotary machinery. A threshold denoising method is used to promote the sparsity of roller bearing and a perfect reconstruction is achieved, which provides a new insight for signal storage and transmission. Furthermore, a fault classification based on compressive sensing is developed in this work without designing a classifier. Compared to other methods of classification, the success ratio is much higher. In addition, a compressed fault detection strategy is proposed to directly detect the fault features from limited samples, which can increase the efficiency of fault diagnosis. Reconstruction and detection may proceed simultaneously without complete recovery and significantly improving detection efficiency is validated by simulations and experiments. The strategy of compressed detection provides a new insight to condition monitoring of rotary machinery, making it possible to largely reduce the data sets while preserving useful information for monitoring. However, there are still lots of un-solved problems still remain for future investigations. Improvements in elimination of more redundant information and preservation of more useful samples will be the focus of our future work regarding compression strategy.

References

1. McFadden P.D., Smith J.D., "Vibration monitoring of rolling element bearings by the high-frequency resonance technique - a review," Tribology international, 1984, 17(1):3–10.
2. Tandon N., Choudhury A., "Review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings", Tribology International, 1999, **32**(8): 469–480.
3. Heng R.B., Nor, M.J., "Statistical analysis of sound and vibration signals for monitoring rolling element bearing condition", Applied Acoustics, 1998, **53**(1):211–226.
4. Junsheng C., Dejie Y., Yu Y., "The application of energy operator demodulation approach based on EMD in machinery fault diagnosis", Mechanical Systems and Signal Processing, 2007, **21**(2):668–677.
5. Ming A.B., Qin Z.Y., Zhang W., Chu, F.L., "Spectrum auto-correlation analysis and its application to fault diagnosis of rolling element bearings", Mechanical Systems and Signal Processing, 2013, **41**(1):141–154.
6. Yu D., Cheng J., Yang Y., "Application of EMD method and Hilbert spectrum to the fault diagnosis of roller bearings", Mechanical Systems and Signal Processing, 2005, **19**(2): 259–270.
7. Rai V.K., Mohanty A.R., "Bearing fault diagnosis using FFT of intrinsic mode functions in Hilbert–Huang transform", Mechanical Systems and Signal Processing, 2007, **21**(6): 2607–2615.
8. Lei Y., Li N., Lin J., Wang S., "Fault diagnosis of rotating machinery based on an adaptive ensemble empirical mode decomposition", Sensors 2013, **13**(12):16950–16964.
9. Yan R., Gao R.X., Chen X., "Wavelets for fault diagnosis of rotary machines: a review with applications", Signal Processing, 2014, **96**(A):1–15.
10. Wang H.Q., Hou W., Tang G., Yuan H.F., "Fault detection enhancement in rolling element bearings via peak-based multi-scale decomposition and envelope demodulation", Mathematical Problems in Engineering, 2014, Article ID 329458.
11. Han J., Kamber M., Pei J., Data Mining Concepts and Techniques, Beijing: Higher Education Pess & Morgan Kaufmann Publishers, 2002.
12. Jerri A.J., "The Shannon sampling theorem – its various extensions and applications: a tutorial review," Proceedings of the IEEE, 1977, **65**(11):1565–1596.
13. Candè E.J., Wakin M.B., "An introduction to compressive sampling," IEEE Signal Processing Magazine, 2008, **25**(2):21–30.
14. Baraniuk R.G., "Compressive sensing," IEEE Signal Processing Magazine, 2007, **24**(4): 118–121.
15. Lustig M., Donoho D.L, Santos J.M., et.al, " Compressed sensing MRI," IEEE Signal Processing Magazine, 2008, **25**(2): 72–82.
16. Tang G., Seismic Data Reconstruction and Denoising based on Compressive Sensing and Sparse Representation, Tsinghua University, Beijing, China, 2010.
17. Davenport M.A., Wakin M.B., Baraniuk R.G., "Detection and estimation with compressive measurements," Tech. Rep, Houston: Rice ECE Department, 2006, pp. 3–13.
18. Haupt J., Nowak R., "Compressed sampling for signal detection," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, Hawaii, April 2007, pp. 1509–1512.
19. Duarte M.F., Davenport M.A., Wakin M.B., et.al, "Sparse signal detection from incoherent projections," Acoustics, Speech and Signal Processing, ICASSP 2006 Proceedings. 2006 IEEE International Conference on. IEEE, 3: III-III.
20. Meng J., Li H., Han Z., "Sparse event detection in wireless sensor networks using compressive sensing", 43rd Annual Conference on IEEE Information Sciences and Systems (CISS 2009), 181–185.
21. Bao Y., Beck J.L., Li H., "Compressive sampling for accelerometer signals in structural health monitoring", Structural Health Monitoring, 2011, **10**(3):235–246.

22. Chen X.F., Du Z.H., Li J.M., Li X., "Compressed sensing based on dictionary learning for extracting impulse components", Signal Processing, 2014, **96**:94–109.
23. Chen X.F., Du Z.H., Li J.M., Li X., "Compressed sensing based on dictionary learning for extracting impulse components", Signal Processing, 2014, **96**:94–109.
24. Zhang X.P., Hu N.Q., Cheng Z.A, "Bearing fault detection method base on compressed sensing". In Engineering Asset Management-Systems, Professional Practices and Certification; Springer International Publishing: New York, NY, USA, 2015; pp. 789–798.
25. Tang G., Yang Q., Wang H.Q., Luo G.G., Ma J.W., "Sparse classification of rotating machinery faults based on compressive sensing strategy", Mechatronics, 2015, **31**:60–67.
26. Tang G., Hou W., Wang H., Luo G.G., Ma J.W., "Compressive sensing of roller bearing faults via harmonic detection from under-sampled vibration signals," Sensors 2015, **15** (10):25648–25662.
27. Tang G., Luo G.G., Ke Y.L., Yang Q., Wang, H.Q., "Compressive sensing: a new insight to signal processing for condition monitoring and fault diagnosis", 28th International Congress of Condition Monitoring and Diagnostic Engineering, 2015, Buenos Aires, Argentina.
28. Davenport M.A., Wakin M.B., Baraniuk R.G., "Detection and estimation with compressive measurements", Dept. of ECE, Rice University, Tech. Rep., 2006.
29. Haupt J., Nowak R., "Compressed sampling for signal detection", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, Hawaii, 2007, 1509–1512.
30. Duarte M.F., Davenport M., Wakin M.B., Baraniuk R.G., "Sparse signal detection from incoherent projections", In Acoustics, Speech and Signal Processing. ICASSP 2006 Proceedings, Vol. 3, pp. III-III.
31. Shannon C.E., "Communication in the presence of noise", Proceedings of the IRE, 1949, 10–21.
32. Mallat S.G., A Wavelet Tour of Signal Processing: The Sparse Way, Academic Press, 2008.
33. Donoho D.L., "Compressed sensing", IEEE Transactions on Information Theory, 2006, 1289–1306.
34. Mallat S.G., Zhang Z., "Matching pursuits with time-frequency dictionaries", IEEE Transactions on Signal Processing, 1993, 3397–3415.
35. Pati Y.C., Rezaiifar R., Krishnaprasad P.S., "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition", Signals, Systems and Computers, 1993 Conference Record of the Twenty-Seventh Asilomar Conference, 1993.
36. Chen S.S., Donoho D.L., "Atomic decomposition by basis pursuit", SIAM Review, 2001, 129–159.
37. Gorodnitsky I.F., Rao B.D., "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," IEEE Transactions on Signal Processing, 1997, 600–616.

Sparse Representation of the Transients in Mechanical Signals

Zhongkui Zhu, Wei Fan, Gaigai Cai, Weiguo Huang and Juanjuan Shi

Abstract This chapter focuses on the sparse representation of the transients in mechanical signals. Sparse representation means that the signal can be represented by an optimal linear combination of atoms by a specialized over-complete dictionary, leading to the sparsity of representation coefficients. Signal sparse representation consists of two main aspects, i.e., dictionary construction and optimization solution. This chapter also presents the applications of sparse representation, mainly in mechanical fault feature detection, such as fault detection of rolling bearings, gearboxes and compound bearing faults.

1 Introduction

Traditional signal representation methods often express a signal as a linear combination of orthogonal atoms which compose a complete dictionary, thus a large number of representation coefficients are required to recover the signal because of the characteristics of the atoms. Instead, by a specialized over-complete dictionary, the signal can be represented by an optimal linear combination of atoms, leading to the sparsity of representation coefficients. Sparse representation has been proven to be one of the powerful tools in signal processing, image processing, computer vision and pattern recognition [1–3]. Recently, much work has been done to introduce sparse representation theory to fault feature extraction from mechanical vibration signals [4–6].

The main purpose of fault diagnosis is to ensure the availability, reliability and operational safety of the equipment. Fault feature extraction which allows one to distinguish the faulty condition from the normal condition is one of the important tasks in fault diagnosis. When there's a defect occurring on the rotating elements (e.g. gear or bearings), it will interact with another element and then produce a

Z. Zhu (✉) · W. Fan · G. Cai · W. Huang · J. Shi
School of Urban Rail Transportation, Soochow University,
Suzhou, People's Republic of China
e-mail: zkzhu@ustc.edu

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors,
Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_9

series of impulses in the vibration signal. Under constant speed operating condition, the vibration responses compose of periodic impulses. These transients display similarly in terms of waveform morphology in the time domain, thus have sparse features. Due to such transient and sparse properties, the fault feature extraction task can be transferred to the task of fault feature sparse representation.

Signal sparse representation consists of two main aspects, i.e., dictionary construction and optimization solution. With a suitably constructed dictionary, few atoms in the dictionary can be merged to represent the fault features effectively, while ineffective in representing the noise. Thus, a sparse representation of the fault features can be obtained, and the background noise can be removed at the same time. Moreover, with an efficient algorithm, the representation coefficients can be easily obtained.

This chapter will present an overview of the sparse representation theory. Moreover it will introduce how to deal with the two main aspects of sparse representation for the mechanical vibration signal processing. The application of sparse representation in mechanical fault feature detection will also be explored in this chapter, such as fault detection of rolling bearings, gearboxes and compound bearing faults.

2 Sparse Representation Theory

2.1 Sparse Representation Model

Consider a matrix $A \in R^{N\times M}$ with $N<M$ whose columns are the atoms $\{\boldsymbol{a}_i\}_{i=1}^{M}$, there are N linear independent vectors in this matrix. The matrix $\boldsymbol{A}$ spans an N-dimension Hilbert space. Suppose the measured fault vibration signal can be written as

$$y(t) = x(t) + n(t) \tag{1}$$

where $y(t)$ is the measured vibration signal, $x(t)$ is the fault-induced signal component without noise and $n(t)$ is the noise. Equation (1) can also be written as

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n} \tag{2}$$

$\boldsymbol{x}$ can be represented with an over-complete matrix $\boldsymbol{A}$ as

$$\boldsymbol{x} = \sum_{i=1}^{M} c_i \boldsymbol{a}_i \tag{3}$$

or more compactly $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{c}$, where $\boldsymbol{c}$ is an $M \times 1$ column vector of representation coefficients. If $\boldsymbol{c}$ is not in the span of the columns of $\boldsymbol{A}$, Eq. (3) has no solution;

otherwise, this equation has infinite number of solutions, with the general solution having l free parameters, where l is the difference between the number of variables and the rank.

Among the general solutions, some may perform better than others. In order to narrow this choice to a well-defined solution, additional criteria are needed [7]. Traditional way to achieve this is to employ the regularization $J(\boldsymbol{c})$. Define the general optimization problem:

$$\min_{\boldsymbol{c}} J(\boldsymbol{c}) \quad \text{s.t. } \boldsymbol{A}\boldsymbol{c} = \boldsymbol{x} \tag{4}$$

There are many possible choices for the objective function $J(\boldsymbol{c})$, from which the well-known choice is the $\|\cdot\|_p$, which denotes the l_p-norm

$$\|\boldsymbol{c}\|_p = \left(\sum\nolimits_i |c_i|^p\right)^{\frac{1}{p}} \tag{5}$$

Let $p \to 0$ of the l_p-norm, the l_0-norm can be denoted as

$$\|\boldsymbol{c}\|_0 = \lim_{p\to 0}\left(\sum_i |c_i|^p\right)^{\frac{1}{p}} \tag{6}$$

In reality, engineers also use the definition of l_0-norm below instead

$$\|\boldsymbol{c}\|_0 = \#\{i{:}c_i \neq 0\} \tag{7}$$

which represents the total number of non-zero elements in a vector. In such underdetermined linear systems of Eq. (3), the aim is to find a sparsest coefficient vector $\boldsymbol{c}$ to "explain" the signal $\boldsymbol{x}$. The sparsest solution means the solution which has the fewest non-zero elements, i.e. the lowest l_0-norm, thus leading to the following equation

$$\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_0 \quad \text{s.t. } \boldsymbol{A}\boldsymbol{c} = \boldsymbol{x} \tag{8}$$

Considering the noise component in the measured signal $\boldsymbol{y}$, Eq. (8) can be written as

$$\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_0 \quad \text{s.t.} \quad \|\boldsymbol{A}\boldsymbol{c} - \boldsymbol{y}\|_2^2 \leq \varepsilon \tag{9}$$

Equation (9) is the sparse representation model.

After the construction of the representation model, two main aspects should be taken into consideration: (a) how to construct a suitable dictionary $\boldsymbol{A}$ to ensure the sparsity of the coefficient vector $\boldsymbol{c}$; (b) how to solve the model to obtain the sparse representation vector. These two problems will be elaborated in the following.

2.2 Construction of the Over-Complete Dictionary

Signal representation can be considered as a way to observe and learn the features of the signal from different aspects. Signal processing techniques commonly require more meaningful representations which can capture the useful characteristics of the signal [8]. The measured signal can be sparsely represented over a dictionary, which means few atoms in the dictionary can be merged to form the signal, indicating that only few coefficients are involved to seize the concerned information. Therefore, it is vital to develop a well-constructed dictionary so that a sparse representation of the signal features can be led. To achieve a proper dictionary, an over-complete dictionary has been established and commonly used.

(a) ***Gabor dictionary***

Gabor atoms are proposed by Dennis Gabor in which a family of functions is built from translations and modulations of a generating function. The Gabor atom is defined as

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right) e^{j\xi t} \tag{10}$$

where $g(t) = e^{-\pi t^2}$ is the Gaussian window function, $\gamma = (s, u, \xi)$ are the parameters of the atom, s is the scaling factor, u is the translation factor, ξ is the frequency factor. The function $g_\gamma(t)$ is centered at u and its energy is mostly concentrated in a neighborhood of u whose size is proportional to s.

Due to the good time-frequency resolution, the Gabor dictionary has been commonly used to analyze the EEG signal [9], audio signal [10] and so on. However, since the atoms of Gabor dictionary are frequency-fixed and divide the time-frequency plane by rectangular grid, it is weak in analyzing signals with frequency-converted components.

(b) ***Chirplet dictionary***

A Chirplet atom, which is built from the unit Gaussian window by dilation, translation, frequency and chirp modulation, is defined as

$$g_\gamma(t) = \frac{1}{\sqrt{\sigma}} g\left(\frac{t-u}{\sigma}\right) \exp\left\{j2\pi\left[1 + r\cdot\left(\frac{t-u}{\sigma}\right)\right]\cdot f_c \cdot \left(\frac{t-u}{\sigma}\right)\right\} \tag{11}$$

Equation (11) can also be written as

$$g_\gamma(t) = \frac{1}{\sqrt{\sigma}} g\left(\frac{t-u}{\sigma}\right) \exp\left\{j2\pi\left[f_c \cdot \left(\frac{t-u}{\sigma}\right) + \frac{1}{2}\xi \cdot \left(\frac{t-u}{\sigma}\right)^2\right]\right\} \tag{12}$$

where $\xi = 2rf_c$ is the linear Chirp rate, $\gamma = (\sigma, u, \xi, c)$ is the atom parameters set, σ is the scale operator which controls the width of the function, u is the time center of

the Chirplet function, ξ is the frequency-variant factor, f_c is the frequency center of the Chirplet function.

Different parameter values denote different Chirplet functions, which compose the Chirplet dictionary. As its frequency can change linearly, the Chirplet dictionary are more likely to be used to describe the signal whose frequency linearly varies with the time, such as the radar signal and the sonar signal [11]. However, when dealing with the time varying components, Chirplet becomes less effective and even inaccurate.

(c) ***FM^{m}let dictionary***

To characterize both the signal's time-invariant and time-varying spectral contents, the dilated and translated windowed exponential frequency modulated functions (FM^{m}let) is proposed by Zou et al. [12]. The FM^{m}let atom is defined as

$$g_\gamma(t) = \frac{1}{\sqrt{\sigma}} g\left(\frac{t-u}{\sigma}\right) \exp\left\{j2\pi\left[1 + r\left(\frac{t-u}{\sigma}\right)\right]^m f_c\left(\frac{t-u}{\sigma}\right)\right\} \tag{13}$$

The atom is expressed by the following five parameters: scaling operator σ, time-center u, frequency center f_c, chirp rate r, and FM exponent m.

The FM^{m}let dictionary is more flexible when dealing with the signal whose spectral contents vary nonlinearly with respect to time; therefor it has been used to process earthquake signal [13], ECG signal [14].

2.3 Solution to Sparse Representation Model

After constructing a suitable dictionary, the next important task is to develop a reliable and efficient algorithm to solve the sparse representation model. It is tough to find a straightforward approach to solve Eq. (9). A number of methods have been developed to solve such an equation in recent years. These methods can be classified into two main categories: the greedy algorithms and the convex relaxation techniques. The greedy algorithms include matching pursuit (MP) [15], orthogonal matching pursuit (OMP) [16], regularized orthogonal matching pursuit (ROMP) [17], etc., and the convex relaxation techniques include basis pursuit (BP) [18], basis pursuit denoising (BPD) [19], etc. In the following, we put an emphasis on the MP and BPD algorithm introduction.

(a) ***Matching pursuit***

Originated from Ref. [15], matching pursuit algorithm uses a greedy heuristic to iteratively construct a best decomposition of the original signal. The basic idea of matching pursuit algorithm is that it attempts to represent a signal $\boldsymbol{x}$ from Hilbert space as a weighted sum of atoms $\boldsymbol{\phi}_{\gamma i}$ taken from an over-complete dictionary $\boldsymbol{\phi}$,

Table 1 Procedures of matching pursuit algorithm

a. **Input**: Dictionary $\boldsymbol{\phi}$ and signal $\boldsymbol{x}$
b. **Initialization**: Iterative parameter $i = 1$, redisual $\boldsymbol{r}_0 = \boldsymbol{x}$ and accuracy requirement ε
c. **Calculation**: Select the optimal atom $\boldsymbol{\phi}_{\gamma i}$ by maximizing $\left\langle \boldsymbol{r}_{i-1}, \boldsymbol{\phi}_{\gamma i} \right\rangle^2$
Compute weighting factor $a = \left\langle \boldsymbol{r}_{i-1}, \boldsymbol{\phi}_{\gamma i} \right\rangle$
Obtain residual $\boldsymbol{r}_i = \boldsymbol{r}_{i-1} - a_{\gamma i}\boldsymbol{\phi}_{\gamma i}$
d. **Stopping rule**: If $\|\boldsymbol{r}_i\|_2 < \varepsilon$, stop. Otherwise, go to step c
e. **Output**: Coefficients $a_{\gamma i}$ after i iterations

$$\boldsymbol{x} = \sum_{i=1}^{m} a_{\gamma i}\boldsymbol{\phi}_{\gamma i} + \boldsymbol{r}_m \tag{14}$$

where a is the weighting factor for each atom, γ is the parameter of each atom, $\boldsymbol{r}_m$ is the residual signal. $\boldsymbol{r}_m$ can be obtained by

$$\boldsymbol{r}_i = \boldsymbol{r}_{i-1} - a_{\gamma i}\boldsymbol{\phi}_{\gamma i} \tag{15}$$

when $i = 1$, $\boldsymbol{r}_0 = \boldsymbol{x}$, the weighting factor $a = \left\langle \boldsymbol{r}_{i-1}, \boldsymbol{\phi}_{\gamma i} \right\rangle$ is the inner product of the residual signal and the atom.

Given the fixed over-complete dictionary, the matching pursuit algorithm first finds the atom which has the biggest inner product with the signal, then subtracts the contribution made by that atom from the residual, and repeats the process until the stopping criterion is satisfied. The procedures of matching pursuit are illustrated in Table 1.

(b) ***Basis pursuit denoising***

Naturally, l_0-norm is used to measure the sparsity of the representation coefficients. However, l_0-norm minimization is a nondeterministic polynomial (NP) problem due to its nature of combinational optimization, which is too complex to solve. By replacing the l_0-norm in Eq. (9) with the l_1-norm, an approximate solution can be attained as follows

$$\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_1 \quad \text{subject to} \quad \|\boldsymbol{A}\boldsymbol{c} - \boldsymbol{y}\|_2^2 \leq \varepsilon \tag{16}$$

where $\|\cdot\|_1$ is the l_1-norm defined as $\|\boldsymbol{c}\|_1 = \sum_i |c_i|$. The basis pursuit denoising can be defined using such an Eq. (16). Donoho [19] has proven that under certain conditions, i.e., the solution is sparse enough, the solution to Eq. (16) is equivalent to the Eq. (9).

Equation (16) can also be written as a more general version

$$J(\boldsymbol{c}) = \arg\min_{\boldsymbol{c}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2 + \lambda \|\boldsymbol{c}\|_1 \right\} \tag{17}$$

where $\|\cdot\|_2^2$ is the l_2-norm defined as $\|\boldsymbol{c}\|_2^2 = \sum_i |c_i|^2$, λ is a scalar regularization parameter which balances the tradeoff between the reconstruction error and the sparsity.

There are two terms in the right side of Eq. (17): the data fidelity term $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2$ and the penalty term $\|\boldsymbol{c}\|_1$. To solve Eq. (17), one can optimize either the data fidelity term or the penalty term. The detailed optimization method will be introduced in Sect. 3.

3 Over-Complete Wavelet Basis Dictionary

3.1 General Over-Complete Wavelet Basis Dictionary

The key point of mechanical fault feature extraction is to construct an appropriate over-complete wavelet basis dictionary. As is well known, the more similarity between the wavelet basis and the fault signal, the sparser the representation coefficients are. According to the experience, the choice of wavelet basis may vary from one to another. Generally, the single-side wavelet is often used to construct the basis matrix for bearing fault signature extraction [20]; the double-side wavelets are usually used for faulty gear vibration signal processing [21]; and chirp signal is often used for radar signal analysis [22], etc. From the literature, the most widely used wavelet basis for mechanical fault feature extraction are more likely to be Laplace wavelet and the Morlet wavelet, which are introduced in detail in the following.

(a) ***Laplace wavelet basis***

The Laplace wavelet is a complex, analytic, single-sided damped exponential wavelet. It is firstly constructed by Strang G in 1996 [23]. Since its waveform is similar to the vibration impulse caused by bearing faults, Laplace wavelet is usually selected to construct the over-complete dictionary for bearing fault detection. As one of the most popular non-orthogonal wavelets, the real field of the Laplace wavelet is defined as

$$\psi(f, \zeta, \tau, t) = \psi_\gamma(t) = \begin{cases} A e^{\frac{-\zeta}{\sqrt{1-\zeta^2}} 2\pi f(t-\tau)} \sin 2\pi f(t-\tau) & t \in [\tau, \tau + W_s] \\ 0 & \text{else} \end{cases} \tag{18}$$

where the parameter vector $\gamma = (f, \zeta, \tau)$ determines the wavelet properties. These parameters (f, ζ, τ) denote frequency $f \in \boldsymbol{R}^+$, damping ratio $\zeta \in [0, 1) \subset \boldsymbol{R}^2$, and time index $\tau \in \boldsymbol{R}$, respectively. The coefficient A is an arbitrary scaling factor used to scale each wavelet to unity norm. The range W_s ensures that the wavelet is compactly supported and has nonzero finite length, but the parameter W_s is generally not explicitly expressed.

(b) ***Morlet wavelet basis***

Morlet wavelet is one of the most popular non-orthogonal wavelets, defined in the time domain as a harmonic wave multiplied by a Gaussian time domain window

$$\psi(t) = \exp\left(-\frac{\beta^2 t^2}{2}\right)\cos(\pi t) \tag{19}$$

It is a cosine signal that decays exponentially on both the left and right sides, which makes it very similar to an impulse caused by the gear localized defects at a constant speed in terms of shape. Therefore, Morlet wavelet is often selected to build the over-complete dictionary when extracting the gear fault feature. In order to reduce the complexity, the parametric formulation of Morlet wavelet is given

$$\psi(f, \zeta, \tau, t) = \psi_\gamma(t) = A\mathrm{e}^{\frac{-\zeta}{\sqrt{1-\zeta^2}}[2\pi f(t-\tau)]^2}\cos 2\pi f(t-\tau) \tag{20}$$

where the parameter vector $\gamma = (f, \zeta, \tau)$ also determines the wavelet properties. These parameters (f, ζ, τ) denote frequency $f \in \boldsymbol{R}^+$, damping ratio $\zeta \in [0, 1) \subset \boldsymbol{R}^+$, and time index $\tau \in \boldsymbol{R}$, respectively. The parameter A is used to normalize the wavelet function.

Setting the discrete parameters f, ζ and τ as the subsets of F, Z and T_c respectively, there is

$$\begin{aligned} F &= \{f_1, f_2, \ldots, f_i\} \subset R \\ Z &= \{\zeta_1, \zeta_2, \ldots, \zeta_j\} \subset R^+ \cap [0, 1) \\ T_c &= \{\tau_1, \tau_2, \ldots, \tau_k\} \subset R \end{aligned} \tag{21}$$

With different parameters, dictionary can be constructed using the following equation

$$\Psi = \{\psi_\gamma(t) : \gamma \in F \times Z \times T_c\} = \{\psi(f, \zeta, \tau, t) : f \in F, \zeta \in Z, \tau \in T_c\} \tag{22}$$

Each item in the dictionary is called an atom. In this way, the over-complete dictionary has been constructed systematically.

3.2 Correlation Filtering

If the suitable wavelet basis (Laplace wavelet, Morlet wavelet or others) is already chosen, correlation filtering is applied to identify the optimal wavelet atom with the optimal set of parameters $(\bar{f}, \bar{\zeta}, \bar{\tau})$, which is most similar to the transient impulses caused by a localized fault.

Correlation, measured by inner product operation, is defined to quantify the degree of similarity between the wavelet basis and the original signal. The correlation function c_γ is defined to calculate the correlation degree between the basis $\psi_\gamma(t)$ and the original signal $x(t)$

$$c_\gamma = \cos\theta = \frac{\left|\left\langle \psi_\gamma(t), x(t)\right\rangle\right|}{\left\|\psi_\gamma(t)\right\|^2 \|x(t)\|^2} \tag{23}$$

where θ is the angle between $\psi_\gamma(t)$ and $x(t)$. The smaller the angle is, the more similar the basis $\psi_\gamma(t)$ and the original signal is. Therefore, the optimal wavelet atom with optimal parameters $(\bar{f}, \bar{\zeta}, \bar{\tau})$ can be obtained by maximizing the correlation function c_γ at each time value from the constructed Laplace wavelet or Morlet wavelet dictionary. Peaks of c_γ for a given time value τ can be represented as

$$k_r(\tau) = \max_{f\in\mathbf{F},\zeta\in\mathbf{Z}} c_\gamma = c\left(\bar{f}, \bar{\zeta}, \tau\right) \tag{24}$$

and the time index parameter $\bar{\tau}$ can be calculated by maximizing the coefficient $k_r(\tau)$. With correlation filtering, the optimal parameters $(\bar{f}, \bar{\zeta}, \bar{\tau})$ found effectively, the optimal wavelet atom with these parameters can be constructed.

4 Solution to Representation Coefficients Based on BPDN

4.1 Data Fidelity Optimization

To represent the fault transients by sparse coefficients, the Basis Pursuit Denoising defined by Eq. (17) should be solved. Only after the minimization of objective function in Eq. (17), a sparse representation vector $\boldsymbol{c}$ can be obtained. To minimize $\boldsymbol{J}(\boldsymbol{c})$, an iterative algorithm is introduced. The traditional gradient descent methods, such as iterative shrinkage/thresholding algorithm (ISTA) [24], fast IST algorithm (FISTA) [25] and so on, have the drawback of slow convergence. In order to improve the speed of convergence, Manya has proposed a novel technique termed the split augmented Lagrangian shrinkage algorithm (SALSA) using the Hessian of

the data fidelity term [26]. The algorithm updates the vector $\boldsymbol{c}$ until the optimal solution $\hat{\boldsymbol{c}}$ is gained, so as to minimize the objective function $\boldsymbol{J}(\boldsymbol{c})$.

Considering the unconstrained optimization problem in which the objective function is the summation of two functions, the Eq. (17) can be written as

$$\min_{\boldsymbol{c}}\{f_1(\boldsymbol{c})+f_2(\boldsymbol{c})\} \tag{25}$$

where $f_1(\boldsymbol{c})=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{c}\|_2^2$, $f_2(\boldsymbol{c})=\lambda\|\boldsymbol{c}\|_1$. Then variable splitting is introduced to create a new variable denoted by $\boldsymbol{u}$, to serve as the augment of f_1, under the constraint that $\boldsymbol{u}=\boldsymbol{c}$. This leads to the constrained problem

$$\min_{\boldsymbol{u},\boldsymbol{c}}\{f_1(\boldsymbol{u})+f_2(\boldsymbol{c})\} \quad \text{s.t.} \quad \boldsymbol{u}=\boldsymbol{c} \tag{26}$$

which is obviously equivalent to the unconstrained problem in Eq. (25). This problem can be represented as the so-called augmented Lagrangian problem

$$\min_{z} E(\boldsymbol{z}) \quad \text{s.t.} \quad \boldsymbol{H}\boldsymbol{z}-\boldsymbol{b}=\boldsymbol{0} \tag{27}$$

where $E(\boldsymbol{z})=f_1(\boldsymbol{u})+f_2(\boldsymbol{c})$, $\boldsymbol{z}=\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{c}\end{bmatrix}$, $\boldsymbol{b}=\boldsymbol{0}$, $\boldsymbol{H}=[\boldsymbol{I}\ -\boldsymbol{I}]$. The augmented Lagrangian function for this problem is defined as

$$L(\boldsymbol{z},\lambda,\mu)=E(\boldsymbol{z})+\lambda^T(\boldsymbol{H}\boldsymbol{z}-\boldsymbol{b})+\frac{\mu}{2}\|\boldsymbol{H}\boldsymbol{z}-\boldsymbol{b}\|_2^2 \tag{28}$$

where λ is a vector of Lagrange multipliers and $\mu\geq 0$ is the penalty parameter. The augmented Lagrangian method (ALM) is used to minimize the objective function $L(\boldsymbol{z},\lambda,\mu)$, the following results can be obtained

$$\boldsymbol{z}^{(k+1)}=\arg\min_{z}\left\{E(\boldsymbol{z})+\frac{\mu}{2}\left\|\boldsymbol{H}\boldsymbol{z}-\boldsymbol{d}^{(k)}\right\|_2^2\right\} \tag{29}$$

$$\boldsymbol{d}^{(k+1)}=\boldsymbol{d}^{(k)}-(\boldsymbol{H}\boldsymbol{z}^{(k+1)}-\boldsymbol{b}) \tag{30}$$

where k is the iteration counter. Considering the concrete forms of the function $E(\boldsymbol{z})$, matrix $\boldsymbol{H}$ and the vector $\boldsymbol{b}$, novel results can be written as

$$\begin{aligned}\boldsymbol{u}^{(k+1)}&=\arg\min_{\boldsymbol{u}}\left\{f_1(\boldsymbol{u})+\frac{\mu}{2}\left\|\boldsymbol{u}-\boldsymbol{c}^{(k)}-\boldsymbol{d}^{(k)}\right\|_2^2\right\}\\&=\arg\min_{\boldsymbol{u}}\left\{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{u}\|_2^2+\frac{\mu}{2}\left\|\boldsymbol{u}-\boldsymbol{c}^{(k)}-\boldsymbol{d}^{(k)}\right\|_2^2\right\}\end{aligned} \tag{31}$$

$$\begin{aligned} \boldsymbol{c}^{(k+1)} &= \arg\min_{\boldsymbol{c}} \left\{ f_2(\boldsymbol{c}) + \frac{\mu}{2} \left\| \boldsymbol{u}^{(k+1)} - \boldsymbol{c} - \boldsymbol{d}^{(k)} \right\|_2^2 \right\} \\ &= \arg\min_{c} \left\{ \lambda \|\boldsymbol{c}\|_1 + \frac{\mu}{2} \left\| \boldsymbol{u}^{(k+1)} - \boldsymbol{c} - \boldsymbol{d}^{(k)} \right\|_2^2 \right\} \end{aligned} \tag{32}$$

$$\boldsymbol{d}^{(k+1)} = \boldsymbol{d}^{(k)} - (\boldsymbol{u}^{(k+1)} - \boldsymbol{c}^{(k+1)}) \tag{33}$$

Equation (31) is a strictly convex quadratic function to be minimized, which leads to the solution $\boldsymbol{u}^{(k+1)}$ directly, and the soft threshold facilitates the minimization of Eq. (32), after which the iteration procedure of SALSA can be listed as

$$\boldsymbol{u}^{(k+1)} = \left(\boldsymbol{A}^H \boldsymbol{A} + \mu \boldsymbol{I}\right)^{-1} \left(\boldsymbol{A}^H \boldsymbol{y} + \mu(\boldsymbol{c}^k + \boldsymbol{d}^k)\right) \tag{34}$$

$$\boldsymbol{c}^{(k+1)} = soft\left(\boldsymbol{u}^{(k+1)} - \boldsymbol{d}^k, \frac{\lambda}{\mu}\right) \tag{35}$$

$$\boldsymbol{d}^{(k+1)} = \boldsymbol{d}^{(k)} - \boldsymbol{u}^{(k+1)} + \boldsymbol{c}^{(k+1)} \tag{36}$$

By the iterative numerical algorithm SALSA, the optimal sparse solution $\hat{\boldsymbol{c}}$ can be obtained eventually. With the sparse solution $\hat{\boldsymbol{c}}$, the reconstructed $\hat{\boldsymbol{x}}$ can be represented as $\hat{\boldsymbol{x}} = \boldsymbol{A}\hat{\boldsymbol{c}}$. There are successive periodic non-zero coefficients in $\hat{\boldsymbol{c}}$, which represents the transients in the original signal.

4.2 Penalty Optimization

Unlike SALSA, Majorization Minimization (MM) algorithm mainly focuses on the penalty optimization to solve the Eq. (17). Based on non-quadratic majorization, the MM algorithm utilizes a sequence of simpler convex optimization problems to replace the original ill-posed inverse problems and yet is an effective and a widely applicable method [27].

The function $\boldsymbol{J}$(c) can be easily minimized suppose it is quadratic. The MM algorithm utilizes this characteristic by solving a series of simpler minimization problems

$$c_{k+1} = \arg\min_{c} G_k(c) \tag{37}$$

where k is the iteration counter, k = 1, 2, 3, …. The MM algorithm requires that each function $G_k(c)$ should be a majorizer (upper bound) of $\boldsymbol{J}$(c) and it coincides with $\boldsymbol{J}$(c) at $c = c_k$. That is

$$\begin{aligned} &\forall c, G_k(c) \geq J(c) \\ &G_k(c_k) = J(c_k) \end{aligned} \tag{38}$$

The majorizer should be chosen so as to be easier to minimize. Considering the data fidelity term in the cost function (17) is strictly quadratic, we simply need to majorize the penalty term.

We mark the penalty $\|c\|_1$ as $\Psi(c)$, term $|c|$ as $\phi(c)$. Hence, $Y(c) = \sum_{n=1}^{N} |c(n)| = \sum_{n=1}^{N} f(c(n))$. $\phi(c)$ is an absolute value function and thus is non-differentiable and non-strictly convex, which makes the Eq. (17) difficult to solve. According to the MM algorithm shown in Eq. (38), a quadratic function $g(c)$ can be found to majorize $\phi(c)$ of the general form

$$g(c) = mc^2 + nc + b \tag{39}$$

where the parameters m, n and b are constants, the majorizer $g(c)$ should be the upper bound for $\phi(c)$ that coincides with $\phi(c)$ at a specified point c_k. For this quadratic majorizer, conditions in Eq. (38) are equivalent to

$$\begin{aligned} g(c_k) &= \phi(c_k) \\ g'(c_k) &= \phi'(c_k) \end{aligned} \tag{40}$$

Solving for m and b gives $m = (\phi'(c_k)/2c_k) - (n/2c_k)$, $b = \phi(c_k) - (c_k/2)\phi'(c_k) - (n/2)c_k$, thus leading to the majorizer $g(c)$ in Eq. (39) given by

$$g(c) = \left(\frac{\phi'(c_k)}{2c_k} - \frac{n}{2c_k}\right)c^2 + nc + \left(\phi(c_k) - \frac{c_k}{2}\phi'(c_k) - \frac{n}{2}c_k\right) \tag{41}$$

Considering a special condition of function $g(c)$, we set the unknown parameter $n = 0$; then the parameter m and b become $m = (\phi'(c_k)/2c_k)$, $b = \phi(c_k) - (c_k/2)\phi'(c_k)$, thus $g(c)$ turns out to be:

$$g(c) = \frac{\phi'(c_k)}{2c_k}c^2 + \phi(c_k) - \frac{c_k}{2}\phi'(c_k) \tag{42}$$

Taking the concrete form of $\phi(c)$ into consideration, the function $g(c)$ can be written in a matrix format as:

$$G_k(c) = \frac{1}{2}\|y - Ac\|_2^2 + \lambda\left(\frac{1}{2}c^*\Lambda_k^{-1}c + \frac{1}{2}\|c_k\|_1\right) \tag{43}$$

where Λ_k denotes the diagonal matrix with vector $|c_k|$ along its diagonal. Then, the MM updates (37) for c_k as:

$$c_{k+1} = \arg\min_{c}\left[\frac{1}{2}\|y - Ac\|_2^2 + \lambda\left(\frac{1}{2}c^*\Lambda_k^{-1}c + \frac{1}{2}\|c_k\|_1\right)\right] \tag{44}$$

The last term in Eq. (44) can be omitted because it does not depend on c; thus a new update equation also called cost function is transformed into:

$$c_{k+1} = \arg\min_{c}\frac{1}{2}\|y - Ac\|_2^2 + \frac{\lambda}{2}c^*\Lambda_k^{-1}c \tag{45}$$

Equation (45) is quadratic in terms of c, so the solution to this problem can be written explicitly using linear algebra as:

$$c_{k+1} = \left(A^*A + \lambda\Lambda_k^{-1}\right)^{-1}A^*y \tag{46}$$

Taking the sparsity of c into consideration, the elements of Λ_k^{-1} would go towards infinity with the iterative procedure going on. To avoid this problem, the matrix inverse lemma is introduced. After that, the update equation can be expressed as:

$$c_{k+1} = \frac{1}{\lambda}\Lambda_k\left[A^*y - A^*(A\Lambda_kA^* + \lambda I)^{-1}A\Lambda_kA^*y\right] \tag{47}$$

By the iterative procedure in Eq. (47) of MM algorithm, the optimal sparse solution $\hat{\boldsymbol{c}}$, which is used to represent the fault feature, can be found. With the sparse solution $\hat{\boldsymbol{c}}$, the reconstructed signal can be represented as $\hat{x} = \boldsymbol{A}\hat{\boldsymbol{c}}$.

5 Applications

5.1 Application in Gearbox Transient Feature Extraction

To verify the effectiveness of the proposed methods for the gearbox fault diagnosis, the experimental data were acquired from an automobile transmission gearbox which has five forward speeds and one backward speed. The structure of the gearbox is shown in Fig. 1. During the test, a broken-tooth fault occurred on the driving gear of the third speed. The vibration signal was acquired by an accelerometer mounted on the outer case of the gearbox when it was loaded with the third speed gearbox.

For a gear transmission, the meshing frequency f_m is calculated by

$$f_m = \frac{nz}{60i} \tag{48}$$

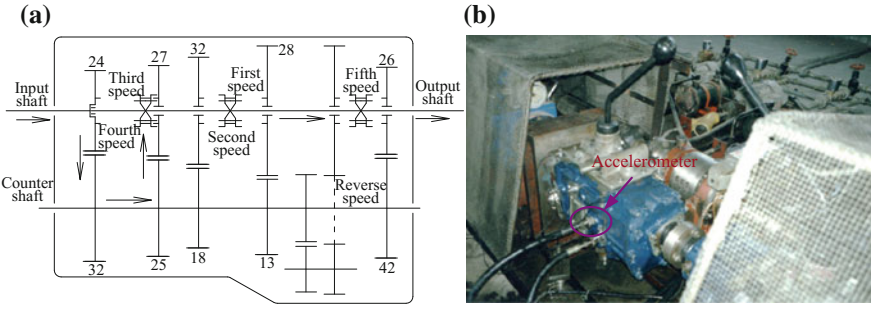


Fig. 1 The automobile transmission gearbox: **a** the structure of gearbox; and **b** gearbox setup

Table 2 Working parameters of the third speed gears

	Number of teeth	Rotating period (ms)	Rotating frequency (Hz)	Meshing frequency (Hz)
Driving gear	25	50.00	20	500
Driven gear	27	54.00	18.5	

where z is the number of the gear teeth, n is the rotating speed of the input shaft in rpm, and i is the transmission ratio. In this test, n was set as 1600 ± 16 rpm, generally we use 1600 rpm as the speed of the input shaft. Then the meshing frequency of the third speed is calculated to be 500 Hz. The sampling frequency was set as 3000 Hz. The working parameters are shown in Table 2.

(a) ***Sparsity-based fault feature extraction by optimizing data fidelity term***

A measured vibration signal with a length of 900 samples and its Fourier spectrum are shown in Fig. 2a, b. The fault feature of the gearbox vibration signal cannot be identified from Fig. 2a. From Fig. 2b, the main frequency component can be identified as 500 Hz, which is in fact the meshing frequency.

Considering the Morlet wavelet is similar to the impulse caused by the gear localized defect, it is selected as the atom to construct the over-complete dictionary. After constructing the dictionary, the sparse representation model can be established. Then the SALSA algorithm can be applied to solve the sparse representation model. Figure 3 presents the analysis result of the vibration signal obtained by the proposed method. Figure 3a displays the optimal wavelet atom based on correlation filtering. The related parameters are $\bar{f} = 272$ Hz, $\bar{\zeta} = 0.0074$ and $\bar{\tau} = 0.0633$ s. The first N elements of the representation coefficient vector $\hat{\mathbf{c}}$ are given in Fig. 3b. 3σ is used as the threshold to filter away the small values to extract the principle components, and then the final estimated vector $\hat{\mathbf{c}}'$ is illustrated in Fig. 3c. The cyclic period $\hat{T} = 50.00$ ms can be easily identified in Fig. 3c, which is consistent with the theoretical value $T = 50.00$ ms. The impulse time can also be identified from Fig. 3c. The parameter values used in this case are listed in Table 3.

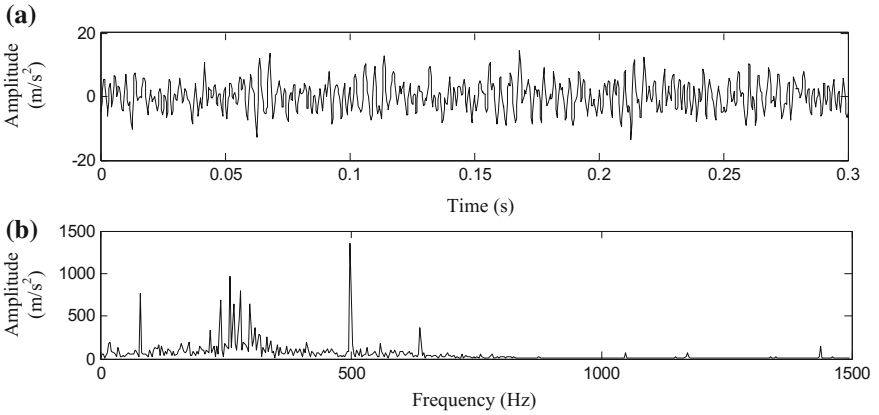


Fig. 2 **a** The measured vibration signal and **b** its Fourier spectrum

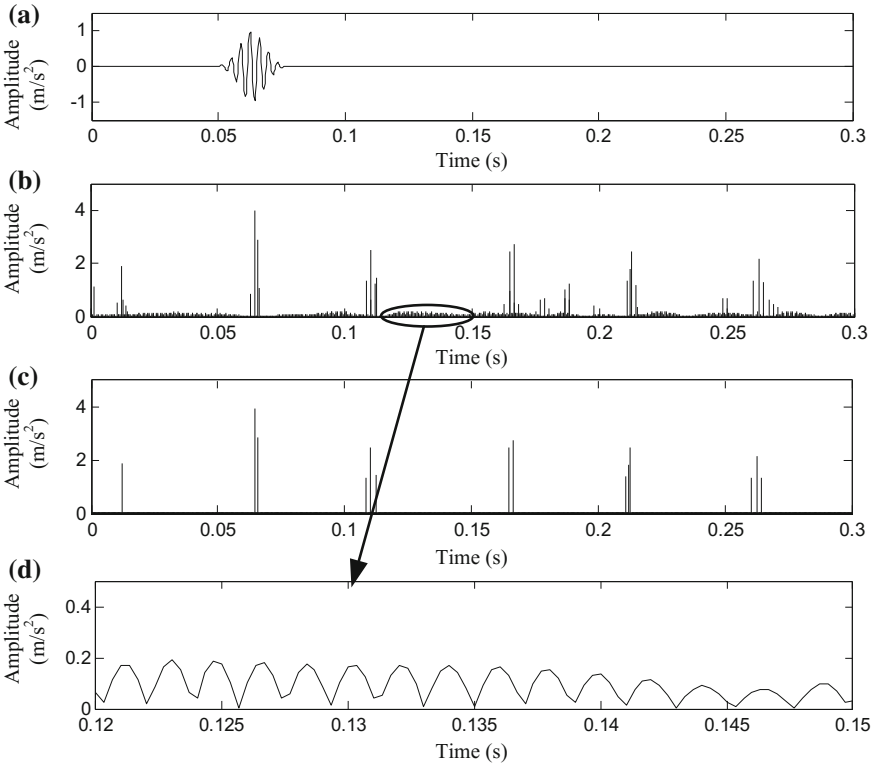


Fig. 3 The representation results of the vibration signal by optimizing the data fidelity term: **a** the optimal wavelet atom; **b** the representation coefficients; **c** the filtered sparse coefficients; and **d** the meshing period

Table 3 Conclusion of all parameters of transient components in the vibration signal of faulty gearbox

Parameters of wavelet basis	$\bar{f} = 272$ **Hz**			$\bar{\zeta} = 0.0074$		
Impulse time (ms)	12.333	64.333	110.333	166.670	212.670	262.670
Period parameter (ms)	50.00					

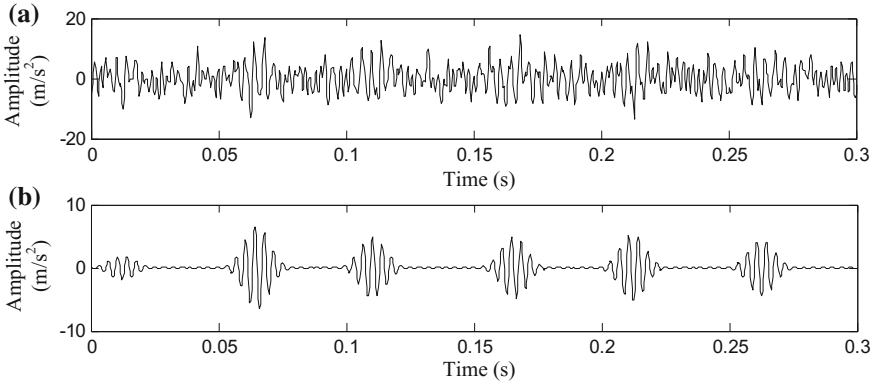


Fig. 4 The comparison between **a** the original vibration signal and **b** the reconstructed signal

The removed small coefficients representing the meshing frequency are shown in Fig. 3d. The meshing period $T_0 = 0.002$ s can be observed in Fig. 3d, indicating that the meshing frequency is 500 Hz. This is consistent with the theoretical meshing frequency 500 Hz. As a result, it is proved that the proposed method is effective in identifying the impulse occurrence time and the period parameter.

A comparison between the reconstructed impulse responses and the original signal is presented in Fig. 4. The interval period between the transients is consistent with the rotating period of the third speed gears. Hence, it indicates that there is a localized fault in the third speed gear of the gearbox. After overhaul, it has been found that the driving gear of the third speed is broken.

(b) ***Sparsity-based fault feature extraction by optimizing penalty term***

Another vibration signal was measured on the same gearbox with a length of 900 and its frequency spectrum is shown in Fig. 5a, b. From Fig. 5a, the impulse period cannot be identified because of the noise corruption; from Fig. 5b, the frequency of the main component can be identified as 500 Hz.

Firstly, the optimal Morlet wavelet atom is obtained by using the correlation filtering to construct the over-complete dictionary. The sparse representation model can subsequently be built. Then the MM algorithm is applied to solve the sparse representation model by optimizing the penalty term. The associated parameters of the optimal wavelet atom are $\bar{f} = 272$ Hz, $\bar{\zeta} = 0.0074$ and $\bar{\tau} = 0.0633$ s. Figure 6a,

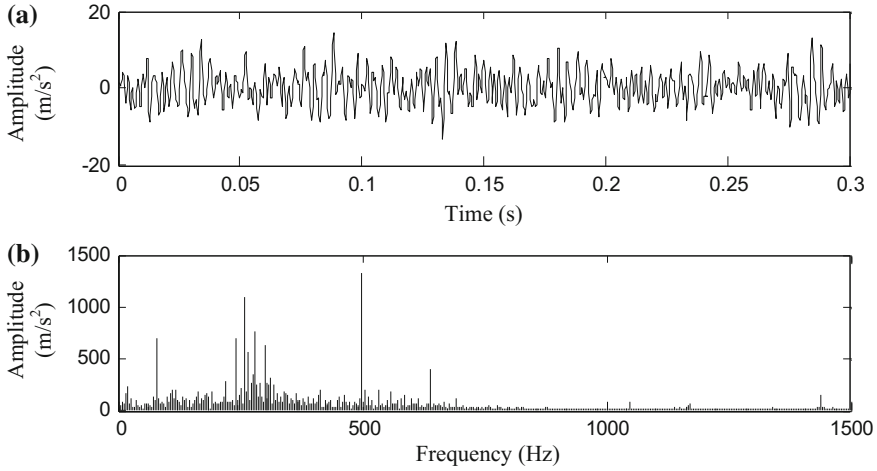


Fig. 5 **a** The measured gearbox defective vibration signal; and **b** its Fourier spectrum

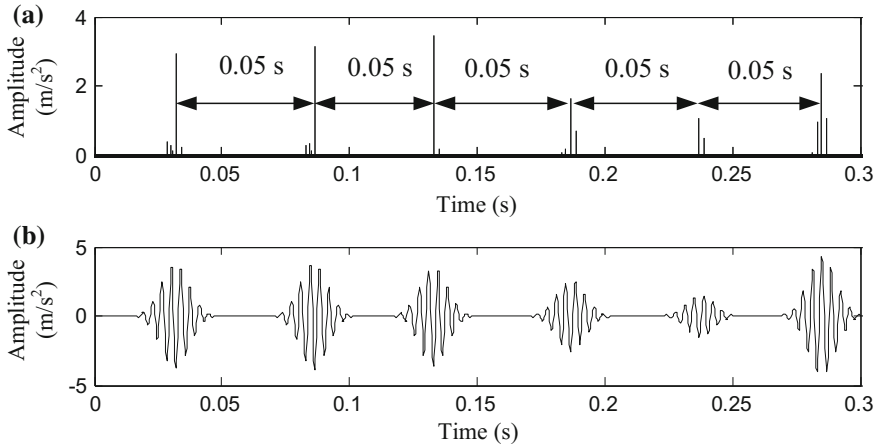


Fig. 6 The analysis results of vibration signal by optimizing the penalty term: **a** sparse coefficients; and **b** the reconstructed signal

obtained by the proposed method, shows the sparse coefficients, which represents a series of periodic impulses. The average time period of these impulses is around 0.0505 s, which is very close to the theoretical value 0.050 s. Figure 6b illustrates the reconstructed signal, whose periodical features represent the localized fault existing in the driving gear of the third speed. The analysis results demonstrate that the proposed transient sparse representation method can extract the transients and reduce the noise effectively, thus the machinery condition can be identified.

5.2 Application in Bearing Transient Feature Extraction

To verify the effectiveness of the proposed method for bearing fault diagnosis, the experimental data were acquired from a test rig, which is shown in Fig. 7. The vibration signal is measured from a rotating machine test rig with the sampling frequency 51.2 kHz. The test rig consists of a driving motor and a shaft, which is driven by the motor and supported by two bearing blocks. The bearing used in this test is NJ208 (TMB) cylindrical roller bearing. Details of the geometry and fault frequencies of this type of bearings can be found in Table 4. In this test, the fault frequency of the outer race, the inner race and the rolling element are 142.8 Hz (7.003 ms), 206.3 Hz (4.847 ms) and 132.6 Hz (7.541 ms), respectively, when the shaft rotates at 1496 RPM.

(c) ***Sparsity-based fault feature extraction by optimizing data fidelity term***

The measured outer race fault vibration signal with 4096 samples and its Fourier spectrum are shown in Fig. 8a, b. The fault feature cannot be seen from Fig. 8. Considering the Laplace wavelet is morphologically similar to the impulse caused by the localized defect in rolling bearing, it is selected as the atom to construct the over-complete dictionary. The sparse representation model can then be founded. To solving the sparse representation model, the analysis results in Fig. 9 can be

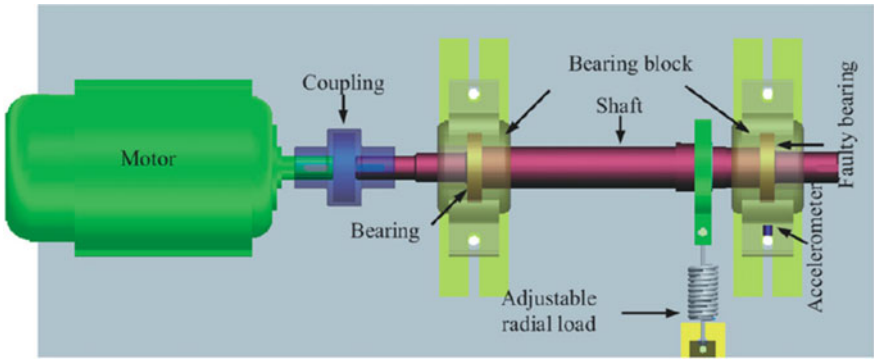


Fig. 7 Rotating machine test rig

Table 4 The geometry and fault frequencies of bearings

Rotating speed of motor (rpm)	1496
Number of rolling elements	14
Inside diameter (mm)	40
Outside diameter (mm)	80
Pitch diameter (mm)	60.5
Roller diameter (mm)	11
Contact angle	0

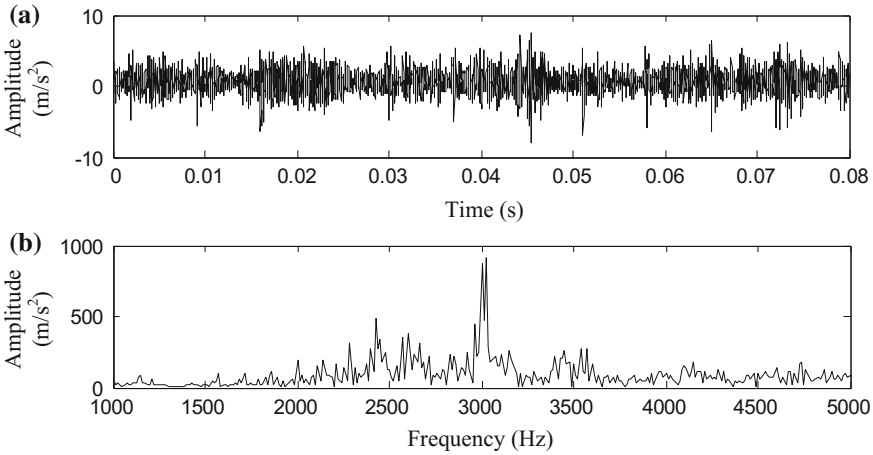


Fig. 8 The measured outer race fault vibration signal. **a** Vibration signal, and **b** Fourier spectrum

obtained by applying the SALSA algorithm. Figure 9a shows the optimal wavelet atom by using correlation filtering. The related parameters are $\bar{f} = 3024$ Hz, $\bar{\zeta} = 0.0890$, $\bar{\tau} = 0.0159$ s. The first N elements of the representation coefficient vector $\hat{\mathbf{c}}$ are given in Fig. 9b. The cyclic period $\hat{T} = 7.01$ ms can be identified in Fig. 9c, which is consistent with the theoretical value $T = 7.00$ ms. The occurrence time of impulse can also be identified in Fig. 9c.

The measured inner race fault vibration signal and its Fourier spectrum are shown in Fig. 10a, b. The fault feature cannot be identified from Fig. 10.

Figure 11 exhibits the analysis result of the inner race vibration signal obtained by the proposed method. The optimal wavelet atom obtained by using correlation filtering is shown in Fig. 11a. The related parameters are $\bar{f} = 6402$ Hz, $\bar{\zeta} = 0.1400$, $\bar{\tau} = 0.0467$ s. The first N elements of the representation coefficient vector $\hat{\mathbf{c}}$ are presented in Fig. 11b. The cyclic period $\hat{T} = 4.85$ ms can be clearly identified in Fig. 11c, which is consistent with the theoretical value $T = 4.85$ ms.

The measured rolling element fault vibration signal and its Fourier spectrum are shown in Fig. 12a, b.

Figure 13 gives the analysis result of the rolling element vibration signal obtained by the proposed method. The optimal wavelet atom obtained by using correlation filtering is shown in Fig. 13a. The related parameters are $\bar{f} = 3024$ Hz, $\bar{\zeta} = 0.089$, $\bar{\tau} = 0.0159$ s. The first N elements of the representation coefficient vector $\hat{\mathbf{c}}$ are displayed in Fig. 13b. The cyclic period $\hat{T} = 7.47$ ms can be identified in Fig. 13c, which is consistent with the theoretical value $T = 7.54$ ms.

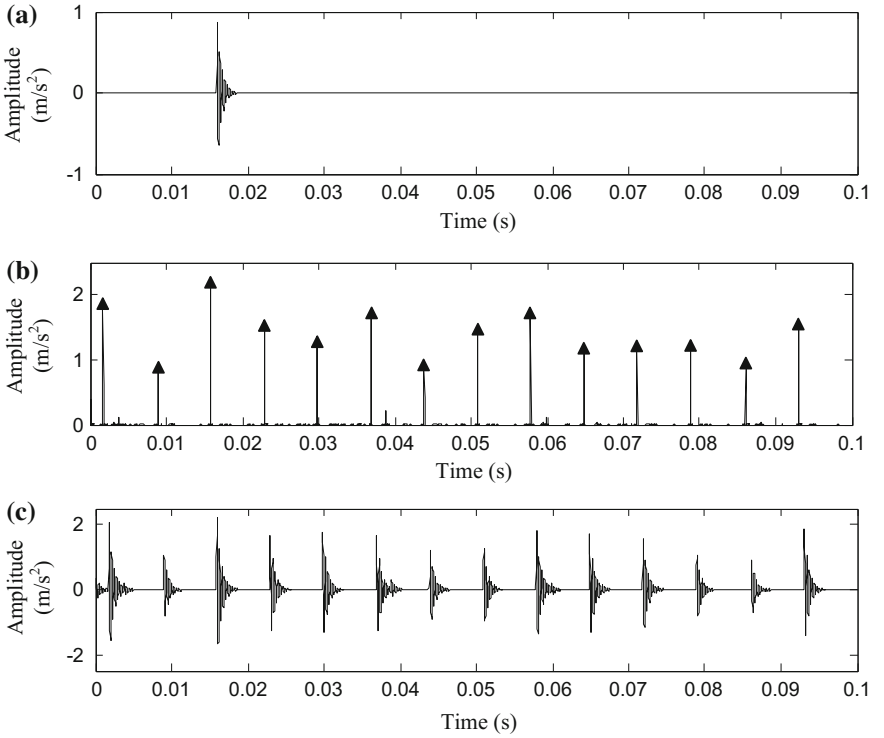


Fig. 9 The analysis results of the outer race fault vibration signal by optimizing the data fidelity term. **a** Optimal Laplace wavelet atom; **b** sparse representation coefficients, and **c** reconstructed signal

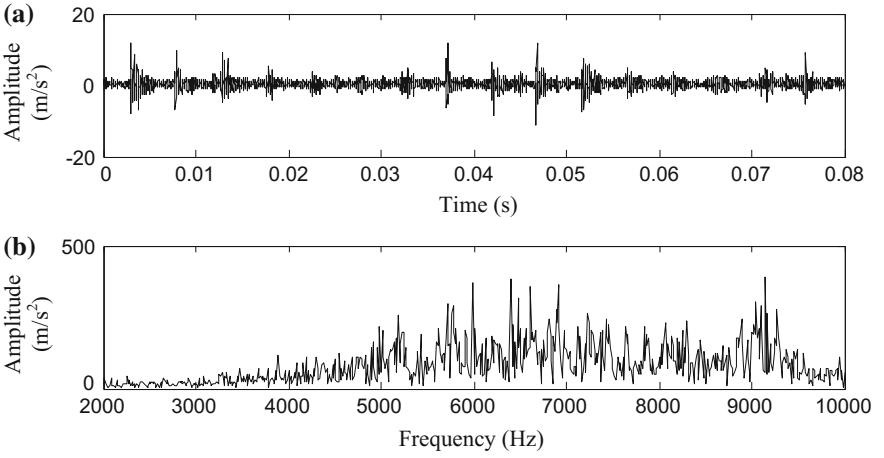


Fig. 10 The measured inner race fault vibration signal. **a** Vibration signal, and **b** Fourier spectrum

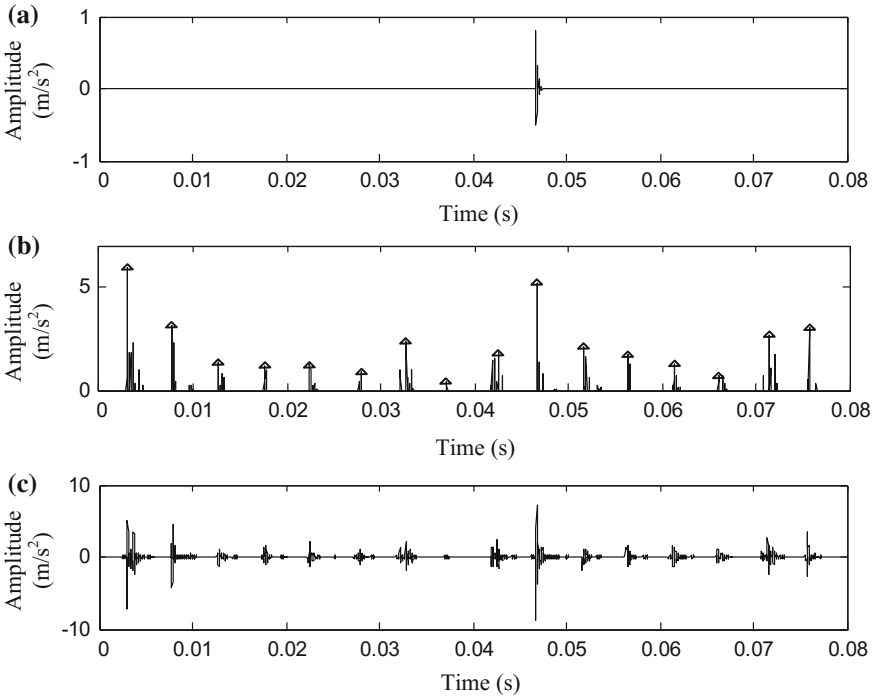


Fig. 11 The analysis results of the inner race fault vibration signal by optimizing the data fidelity term. **a** Optimal Laplace wavelet atom, **b** sparse representation coefficients, **c** reconstructed signal

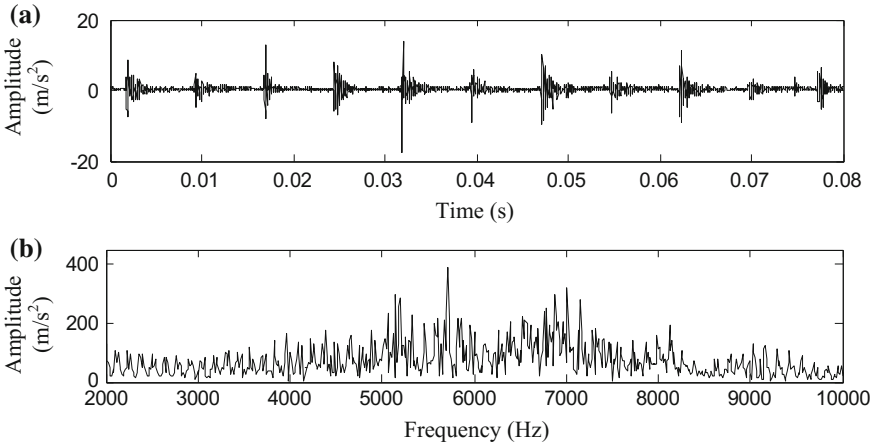


Fig. 12 The measured rolling element fault vibration signal: **a** vibration signal, and **b** Fourier spectrum

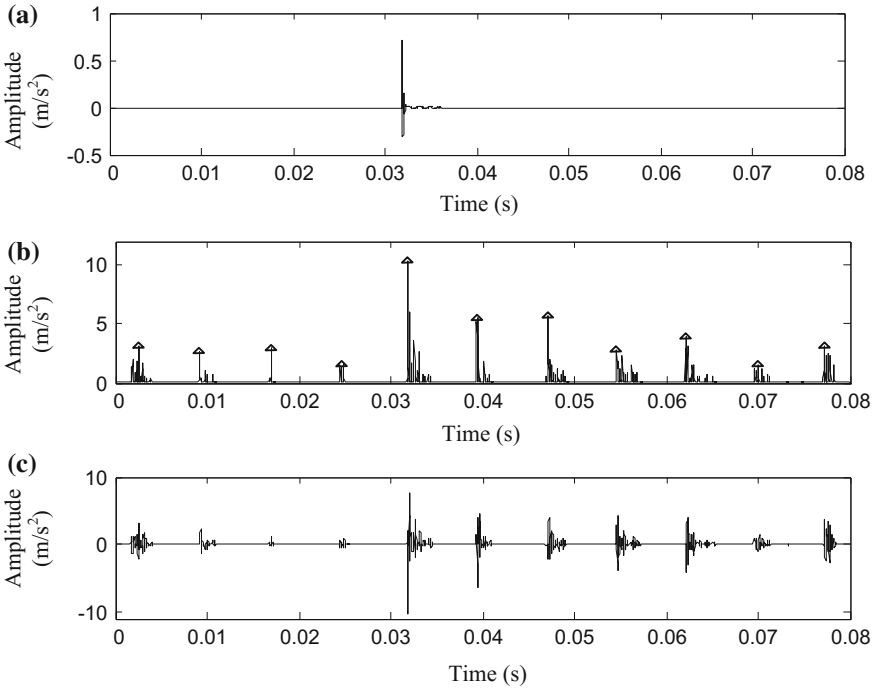


Fig. 13 The analysis results of the rolling element fault vibration signal by optimizing the data fidelity term: **a** Optimal Laplace wavelet atom, **b** sparse representation coefficients, and **c** reconstructed signal

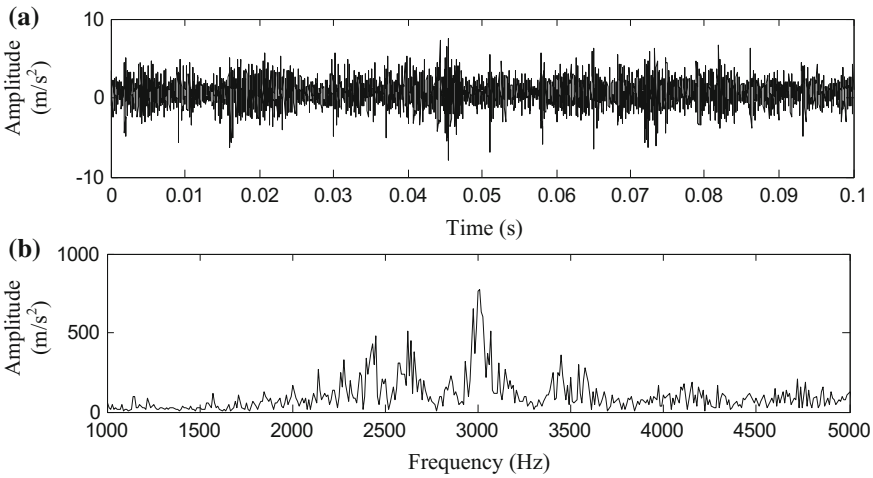


Fig. 14 The measured outer race fault vibration signal: **a** vibration signal, and **b** Fourier spectrum

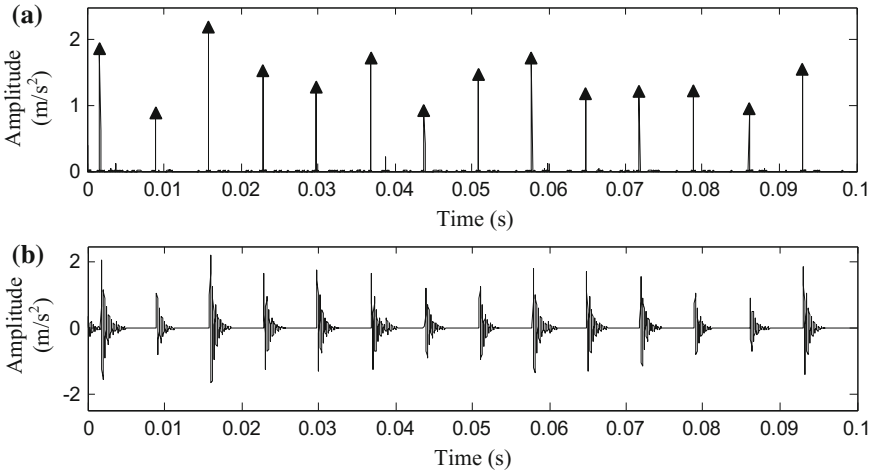


Fig. 15 The analysis results of the outer race fault vibration signal by optimizing penalty term: **a** sparse representation coefficients, and **b** reconstructed signal

(d) ***Sparsity-based fault feature extraction by optimizing penalty term***

Another group of fault signals with a length of 5120 samples is also measured from the same test rig as Fig. 7. The over-complete dictionary is constructed using the Laplace wavelet, from which a signal sparse representation model can be established. Then the MM algorithm is applied to solve the model to obtain the representation coefficients. The measured outer race fault vibration signal and its Fourier spectrum are shown in Fig. 14a, b. No information related to bearing defects can be recognized.

Figure 15 gives the analysis result of the vibration signal obtained by the proposed method. The representation coefficient vector $\hat{\mathbf{c}}$ is given in Fig. 15a, in which the cyclic period $\hat{T} = 7.02$ ms can be discerned. The reconstructed signal is shown in Fig. 15b.

The measured inner race fault vibration signal and its Fourier spectrum are shown in Fig. 16a, b, which cannot easily determine the bearing health condition.

Figure 17 shows the analysis result of the vibration signal obtained by the proposed method. The representation coefficient vector $\hat{\mathbf{c}}$ is given in Fig. 17a, in which the cyclic period $\hat{T} = 4.84$ ms can be discerned. The reconstructed signal is shown in Fig. 17b, yielding the easily-observed impulses.

The measured rolling element fault vibration signal and its Fourier spectrum are shown in Fig. 18a, b. No frequency information associated with rolling elements can be recognized in Fig. 18b.

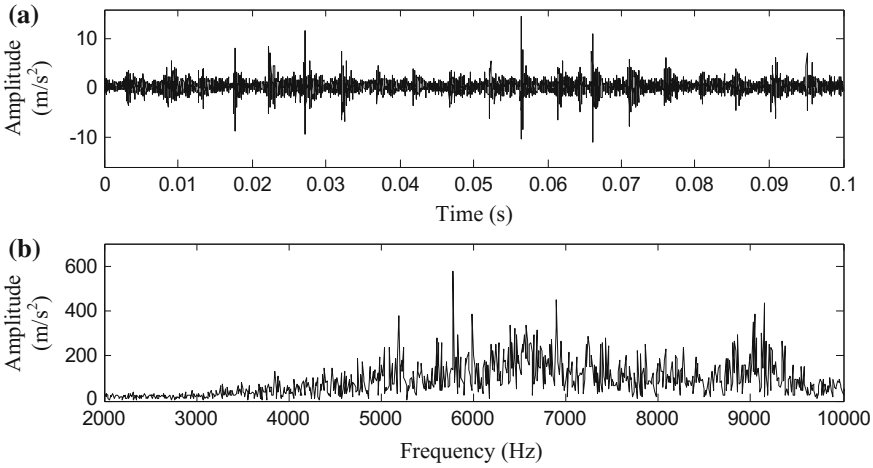


Fig. 16 The measured inner race fault vibration signal: **a** vibration signal, and **b** Fourier spectrum

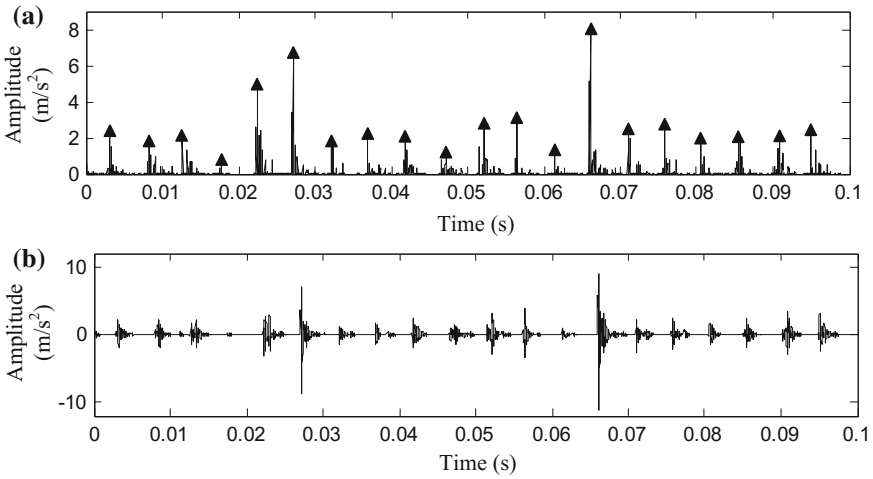


Fig. 17 The analysis results of the inner race fault vibration signal by optimizing penalty term: **a** sparse representation coefficients, and **b** reconstructed signal

Figure 19 exhibits the analysis result of the vibration signal obtained by the proposed method. The representation coefficient vector $\hat{\mathbf{c}}$ is shown in Fig. 19a, where the cyclic period $\hat{T} = 7.51$ ms can be identified. The reconstructed signal shown in Fig. 19b clearly shows the cyclic impulses generated by rolling element defect.

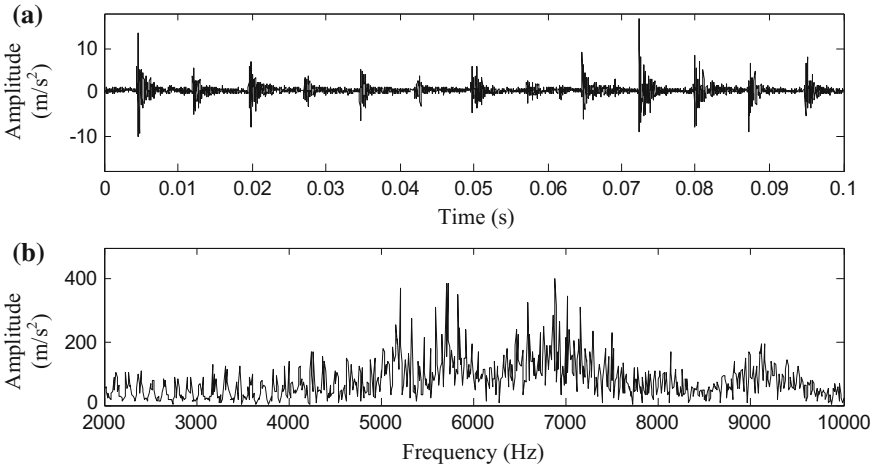


Fig. 18 The measured rolling element fault vibration signal: **a** vibration signal, and **b** Fourier spectrum

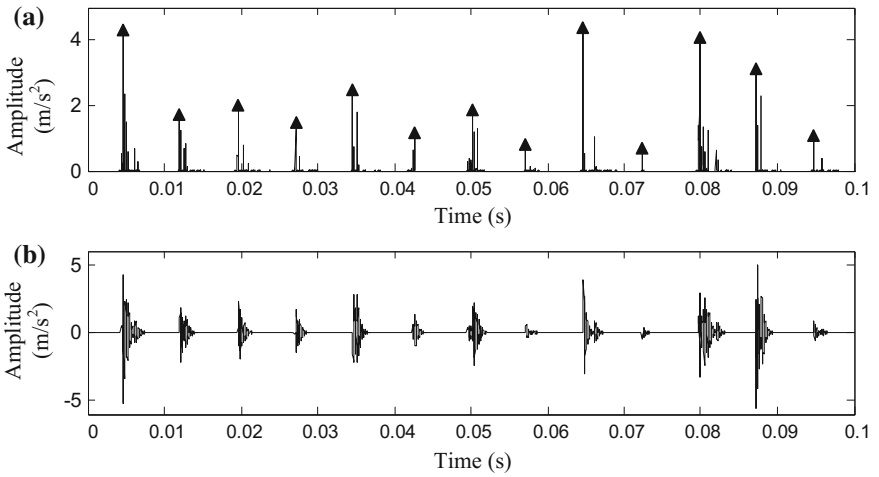


Fig. 19 The analysis results of the rolling element fault vibration signal by optimizing penalty term: **a** sparse representation coefficients, and **b** reconstructed signal

5.3 Application in Compound Fault Feature Extraction

Apart from single fault detection of rotating machinery, compound fault diagnosis also has been gaining more attention in recent years. Taking the compound fault in the gearbox as an example, this section applies the sparse representation method to separating and extracting the compound fault features of gearbox.

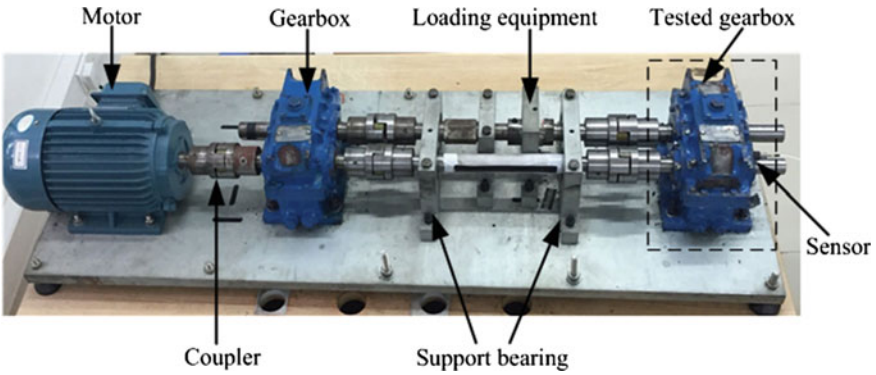
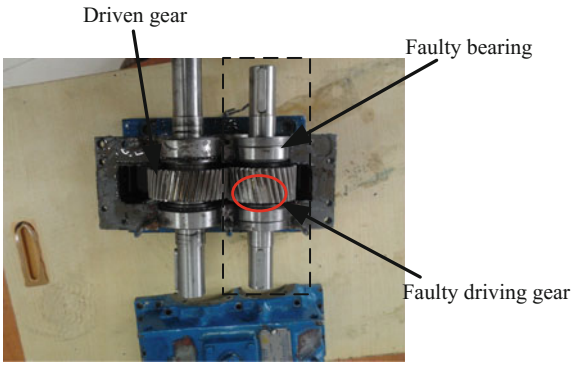


Fig. 20 Experimental gearbox in a test-bed

Fig. 21 Fault components



The test rig, which is a single stage transmission gearbox in a test-bed, is shown in Fig. 20. There are both bearing and gear faults in the gearbox, which is shown in Fig. 21, respectively. The faulty gear is a helical, whose parameters are listed in Table 5. The bearing model in the experiment is 30,205, taper roller bearing, and its geometric parameters are listed in Table 6. With the known parameters, the theoretical fault feature frequency of the bearing can be calculated as 176.18 Hz.

The measured vibration signal with compound faults is shown in Fig. 22, from which the characteristics of each fault cannot be identified clearly. Thus, the sparse representation method is applied to extracting the fault features one by one. In terms of the sequence of the compound fault feature extraction, we take the influence of propagation path of signals into consideration. As the sensor is placed on the bearing end cover, which is closer to the faulty bearing, it is desirable to extract the bearing fault feature at first. Firstly, the iterative algorithm SALSA is selected as the optimization algorithm. Then, the optimal Laplace wavelet, which is effective in bearing fault induced impulse representation and determined using the correlation filtering, is chosen to construct the over-complete dictionary $\boldsymbol{A}_1$ based on the explanation in Sect. 2. The selected Laplace wavelet is shown in Fig. 23a.

Table 5 Working parameters of gears in the tested gearbox

Gear	Number of teeth	Rotating frequency (Hz)	Rotating period (ms)	Meshing frequency (Hz)
Driving gear	34	24.67	41	839
Driven gear	42	19.98	50	

Table 6 Geometry of the tested bearing

Inside diameter (mm)	Outside diameter (mm)	Ball diameter (mm)	Number of rolling elements	Contact angle (°)
30	62	8	17	14

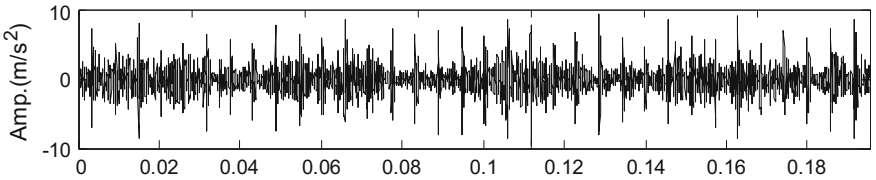


Fig. 22 The measured signal with compound fault of gearbox

Incorporating the dictionary $\boldsymbol{A}_1$ into the iterative procedure of SALSA, the sparse coefficients $\hat{\boldsymbol{c}}_1$ of bearing fault can be obtained in Fig. 23b. The corresponding reconstructed signal illustrated in Fig. 23c can also be obtained by the equation $\hat{x}_1 = \boldsymbol{A}\hat{\boldsymbol{c}}_1$. To acquire the fault characteristics, the envelope spectrum analysis of the reconstructed signal is performed, yielding the result in Fig. 23d. The characteristic frequency of the faulty bearing, 174.1 Hz, can be easily recognized, which is almost identical to the theoretical value 176.18 Hz. Therefore, it can be concluded that the bearing is defective.

As we know, the amplitude of each transient impulse caused by localized bearing fault is represented by the sparse vector $\hat{\boldsymbol{c}}_1$. In order to estimate the real amplitude of bearing fault transients, a constrained optimization strategy is proposed is proposed to estimate the amplitude of each single fault component by introducing the parameter k. The spectrum of the residual fault signal $x - k\hat{x}_1$ is denoted by $F_1(f)$

$$\begin{aligned} &\min\{F_1(f)\} \\ &\text{subject to } k > 0, f = f_{z1} \end{aligned} \tag{49}$$

where x is the original measured signal, f_{z1} is the peak frequency and k is a positive parameter. When $F_1(f)$ is minimized subject to its constraints, it indicates that the bearing fault component in the residual fault signal has been removed to the largest extent. By solving problem in (48), an optimal value k_{opt} is acquired and the estimated bearing fault signal can be obtained by the function $x_1 = k_{opt}\hat{x}_1$.

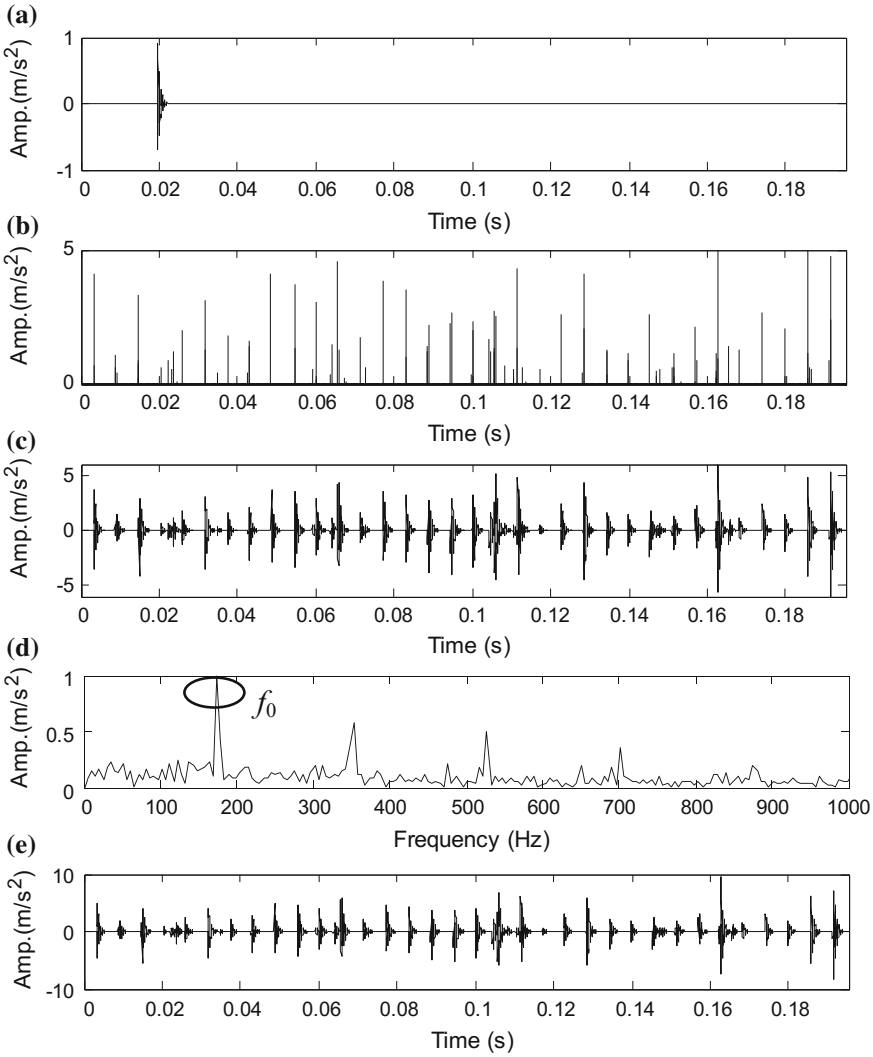


Fig. 23 Results of bearing fault signal: **a** optimal Laplace basis, **b** sparse coefficients, **c** reconstructed signal, **d** the envelope spectrum analysis of the reconstructed signal, and **e** the estimated bearing fault signal

Based on the above description, we can draw that Fig. 23e shows the estimated bearing fault component with $k_{opt} = 1.332$.

Removing the estimated bearing fault signal from the original signal, the residual signal is shown in Fig. 24. Similar to the bearing fault feature extraction, the SALSA is firstly chosen as the optimization algorithm. Then, the optimal Morlet wavelet basis A2 is obtained by correlation filtering, as presented in Fig. 25a.

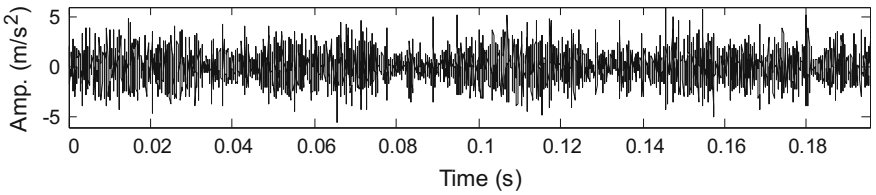


Fig. 24 The residual signal after removing the bearing fault signal

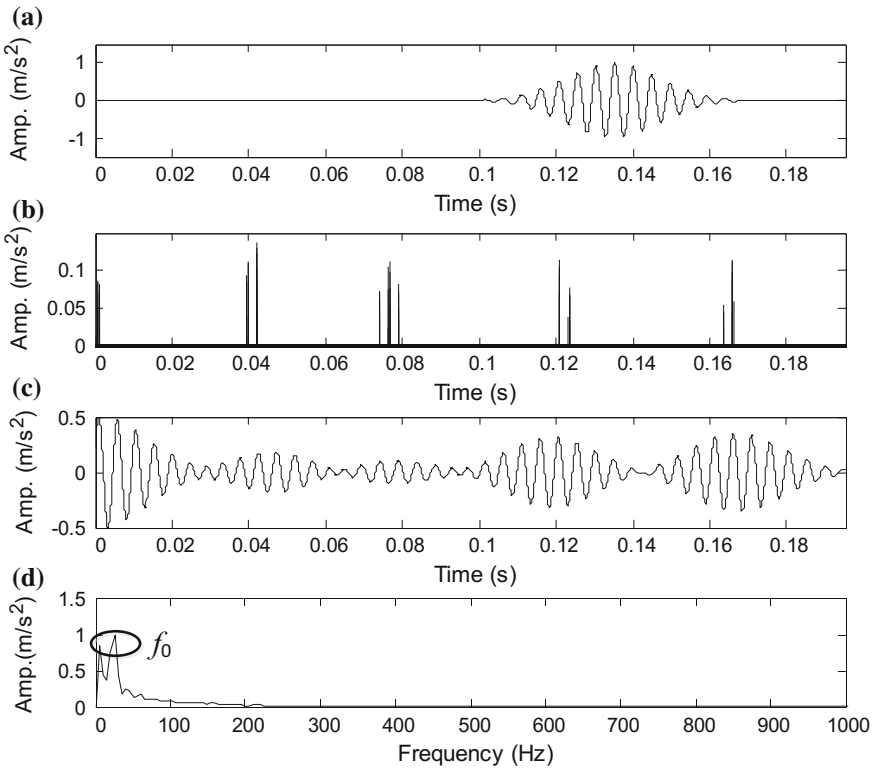


Fig. 25 Results of gear fault signal: **a** optimal Morlet basis, **b** sparse coefficients, **c** reconstructed signal, and **d** the envelope spectrum analysis of the reconstructed signal

With the constructed dictionary A2, the iterative procedure can be implemented to gain the sparse vector $\hat{c}_2$ representing the gear fault feature, as shown in Fig. 25b. Its reconstructed signal is obtained in Fig. 25c. The envelope spectrum analysis of the reconstructed signal is illustrated in Fig. 25d, from which the fault characteristic frequency of gear can be identified, 25.6 Hz, close to the theoretical value 24.67 Hz. The analysis indicates that there is a gear localized fault in the tested gearbox.

6 Discussions

In this chapter, a new transient extraction technique is introduced based on the sparse representation. To be more specific, the sparse representation model and over-complete dictionary are first constructed, and then the model can be solved by optimizing either the data fidelity term or the penalty term. Both are effective in extracting the transients and identifying the periodic parameters. The effectiveness has been demonstrated by the experimental applications. However, some issues about the proposed method still remain to be discussed.

(1) In this chapter, the l_1-norm is used to replace the l_0-norm in the sparse representation model. Another available sparsity measurement method is to use the l_p-norm, leading to the following equation:

$$\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_p^p \quad \text{s.t.} \quad \|\boldsymbol{A}\boldsymbol{c} - \boldsymbol{y}\|_2^2 \leq \varepsilon \tag{50}$$

Choosing $p < 1$ will lead to a sparse solution; however, it will also lead to a non-convex optimization problem. Thus we can use $J(\boldsymbol{c}) = \sum_i \rho(c_i)$ to replace the lp-norm. Actually, any function $J(\boldsymbol{c}) = \sum_i \rho(c_i)$ with $\rho(c_i)$ being symmetric, monotonically non-decreasing, and with a monotonic non-increasing derivative for $c \geq 0$ will lead to the sparsity [7].

(2) Selection of the wavelet basis is one of the key issues for the proposed method due to its influences on the sparsity of the coefficient vector $\boldsymbol{c}$. With the increase of the noise amplitudes, the correlation values decrease sharply and thus leading to an error between the estimated value and the theoretical one. Besides, the empirical knowledge about the gearbox fault and bearing fault is used to construct the over-complete dictionary. Therefore, if the dictionary can learn from the measured signal by adding some rotating component fault features, the algorithms in this chapter will be more powerful in mechanical fault diagnosis.

(3) The strategy of optimal wavelet atom determination and the algorithms of solving the sparse representation model are also vital for a successful sparse representation application to machinery fault feature extraction.

- This chapter employs the correlation filtering for the optimal wavelet atom selection. The disadvantage is that larger interval range and smaller step of the parameter subset $\boldsymbol{\Psi}$, which can increase the accuracy of the result though, would incur excessive computation, thereby decreasing the efficiency of the method. Therefore, the strategy of optimal wavelet basis selection should be further exploited to ensure not only the computational efficiency but also estimation accuracy.
- This chapter utilizes the SALSA and MM algorithm to optimize the BPD problem. However, the more straightforward and simpler, yet effective, algorithms have not been largely explored for the solution of sparse representation model.

References

1. Bruckstein A.M., Donoho D.L., Elad M., "From sparse solutions of systems of equations to sparse modeling of signals and images," SIAM Review, 2009, 51(1): 34–81.
2. Yang J., Wright J., Huang T.S., et al., "Image super-resolution via sparse representation," IEEE Transactions on Image Processing, 2010, 19(11): 2861–2873.
3. Wright J., Ma Y., Mairal J., et al., "Sparse representation for computer vision and pattern recognition," Proceedings of the IEEE, 2010, 98(6): 1031–1044.
4. Liu H., Liu C., Huang Y., "Adaptive feature extraction using sparse coding for machinery fault diagnosis," Mechanical Systems and Signal Processing, 2011, 25(2): 558–574.
5. Peng F., Yu D., Luo J., "Sparse signal decomposition method based on multi-scale Chirplet and its application to the fault diagnosis of gearboxes," Mechanical Systems and Signal Processing, 2011, 25(2): 549–557.
6. Yang H., Mathew J., Ma L., "Fault diagnosis of rolling element bearings using basis pursuit," Mechanical Systems and Signal Processing, 2005, 19(2): 341–356.
7. Elad M., Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing, Springer, Israel, 2010.
8. Rubinstein R., Bruckstein A.M., Elad M., "Dictionaries for sparse representation modeling," Proceedings of the IEEE, 2010, 98(6): 1045–1057.
9. Aviyente S., "Compressed sensing framework for EEG compression," IEEE/SP 14th Workshop on Statistical Signal Processing, 2007: 181–184.
10. Ghoraani B., Krishnan S., "Time–frequency matrix feature extraction and classification of environmental audio signals," IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2197–2209.
11. Yeste-Ojeda O.A., Grajal J., López-Risue G., "Atomic decomposition for radar applications," IEEE Transactions on Aerospace and Electronic Systems, 2008, 44(1): 187–200.
12. Zou H., Dai Q., Wang R., et al., "Parametric TFR via windowed exponential frequency modulated atoms," IEEE Signal Processing Letters, 2001, 8(5): 140–142.
13. Li X., Zhao K., Liu D., et al., "Feature extraction and identification of underground nuclear explosion and natural earthquake based on FMmlet transform and BP neural network," Advances in Neural Networks-ISNN 2004. Springer Berlin Heidelberg, 2004: 925–930.
14. Meng Q.J., Sun N., "A comparison study on Gabor, Chirplet, FMm let atom databases for ECG signal processing," 3rd IEEE International Conference on Bioinformatics and Biomedical Engineering, 2009: 1–3.
15. Mallat S.G., Zhang Z., "Matching pursuits with time-frequency dictionaries," IEEE Transactions on Signal Processing, 1993, 41(12): 3397–3415.
16. Donoho D.L., Tsaig Y., Drori I., et al., "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit," IEEE Transactions on Information Theory, 2012, 58(2): 1094–1121.
17. Needell, D., Vershynin, R.: "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," Found. Comput. Math., 2009, 9(3):317–334.
18. Chen, S.S., Donoho, D.L., Saunders, M.A., "Atomic decomposition by basis pursuit," SIAM J. Sci. Comput., 1998, 20(1), 33–61.
19. Donoho D.L., Huo X., "Uncertainty principles and ideal atomic decomposition," IEEE Transactions on Information Theory, 2001, 47(7): 2845–2862.
20. Feng K., Jiang Z., He W., et al., "Rolling element bearing fault detection based on optimal antisymmetric real Laplace wavelet," Measurement, 2011, 44(9): 1582–1591.
21. Zheng H., Li Z., Chen X., "Gear fault diagnosis based on continuous wavelet transform," Mechanical Systems and Signal Processing, 2002, 16(2): 447–457.
22. Droitcour A., Boric-Lubecke O., Lubecke V., et al., "Range correlation and I/Q performance benefits in single-chip silicon Doppler radars for noncontact cardiopulmonary monitoring," IEEE Transactions on Microwave Theory and Techniques, 2004, 52(3):838–848,
23. Strang G., Nguyen T., Wavelets and Filter Banks, SIAM, 1996.

24. Daubechies I., Defrise M., De Mol C., "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," Commun Pur Appl Math, 2003, 57: 1413–1457.
25. Beck A., Teboulle M., "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," SIAM Journal on Imaging Sciences, 2009, 2(1): 183–202.
26. Afonso M.V., Bioucas-Dias J.M., Figureueiredo M.A.T., "Fast image recovery using variable splitting and constrained optimization," IEEE Transactions on Image Processing, 2010, 19(9): 2345–2356.
27. Hunter D.R., Lange K., "A tutorial on MM algorithms," The American Statistician, 2004, 58 (1): 30–37.

Fault Diagnosis of Rotating Machinery Based on Empirical Mode Decomposition

Yaguo Lei

Abstract Rotating machinery covers a broad range of mechanical equipment in industrial applications. It generally operates under tough working environment and is therefore subject to faults easily. Vibration signals collected in the working process have valuable contributions for the presentation of conditions of the rotating machinery. Consequently, using signal processing techniques, these faults could be detected and diagnosed. Empirical mode decomposition (EMD) is one of the most powerful signal processing techniques and has been widely applied in fault diagnosis of rotating machinery. This chapter attempts to introduce the recent research and development of EMD in fault diagnosis of rotating machinery, including basic concepts and fundamental theories about EMD methods and improved EMD methods. Moreover, the applications of EMD methods and improved EMD methods in fault diagnosis of common and key components of rotating machinery, like rotors, gears and rolling element bearings, are described in details.

1 Introduction

Rotating machinery plays an important role in industrial applications. It generally works under a tough environment. Thus rotating machinery can suffer from failures easily, which may decrease the service performance such as manufacturing quality, operation safety, etc., and even cause the entire mechanical system to break down. With rapid development of science and technology, rotating machinery is becoming larger, more precise and more automatic. Its potential faults become more difficult to be detected. Accordingly, the investigations of rotating machinery fault diagnosis have attracted considerable interests in recent years. Vibration signals collected in the working process have valuable contributions for the presentation of conditions

Y. Lei (✉)
State Key Laboratory for Manufacturing Systems Engineering,
Xi'an Jiaotong University, Xi'an, China
e-mail: yaguolei@mail.xjtu.edu.cn

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors,
Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_10

of the rotating machinery. Consequently, adopting advanced signal processing techniques to reveal fault characteristics is one of the commonly used strategies in fault diagnosis of rotating machinery [1, 2]. Empirical mode decomposition (EMD) is one of the most advanced signal processing techniques [3], which is proposed as an adaptive time-frequency signal processing method to analyze non-stationary and nonlinear signals. It is based on the local characteristic time scales of a signal and could decompose the signal into a set of complete and almost orthogonal components called intrinsic mode functions (IMFs). The IMFs indicate the natural oscillatory mode imbedded in the signal and serve as the basis functions, which are determined by the signal itself, rather than pre-determined kernels. Thus, it is a self-adaptive signal processing technique that is suitable for nonlinear and non-stationary processes. Since EMD was proposed in 1998, it has been widely utilized and extensively studied in a lot of areas, for example, process control [4, 5], modeling [6–8], surface engineering [9], medicine and biology [10], voice recognition [11], system identification [12, 13], etc.

Although EMD largely contributes to the analysis of non-stationary and nonlinear signals, the algorithm itself has some shortcomings [14–16], such as end effects, mode mixing, etc. Aiming at these drawbacks, various theoretical analyses and improved EMD methods have been accomplished [17–22]. In addition, some improved EMD methods have been applied in the diagnosis of early rub-impact faults of rotors [21, 22], crack faults of gears [23, 24], and single or compound faults of locomotive bearings [25, 26].

This chapter attempts to introduce the recent research and development of EMD in fault diagnosis of rotating machinery. In the rest of this chapter, basic concepts and fundamental theories about EMD methods and improved EMD methods will be presented. In addition, the applications of EMD methods and improved EMD methods in fault diagnosis of rotors, gears and rolling element bearings, which are the common and key components of rotating machinery, will be described in details.

2 Empirical Mode Decomposition

2.1 EMD Algorithm

The EMD algorithm was proposed by Huang et al. and could decompose a signal into a set of IMFs [3]. An IMF is a function that should be satisfied with the following two conditions: (1) in the whole data set, the number of extrema and the number of zero-crossings must either equal or differ at most by one, and (2) at any point, the mean value of the envelope defined by local maxima and the envelope defined by the local minima is zero [3]. An IMF represents the natural oscillatory mode embedded in the signal. A typical IMF is shown in Fig. 1.

With the simple assumption that any signal consists of different simple IMFs, the EMD method could decompose a signal into some IMF components, which are determined by the signal itself. Thus, it is a self-adaptive signal processing method. Given a signal $x(t)$, the EMD algorithm can be described as follows.

(1) Initialize: $r_0 = x(t)$, and $i = 1$.
(2) Extract the i-th IMF.

   (a) Initialize: $h_{i(k-1)} = r_i$, $k = 1$.
   (b) Extract the local maxima and minima of $h_{i(k-1)}$.
   (c) Interpolate the local maxima and the minima by cubic spline lines to form upper and lower envelops of $h_{i(k-1)}$.
   (d) Calculate the mean $m_{i(k-1)}$ of the upper and lower envelops of $h_{i(k-1)}$, as shown in Fig. 2.
   (e) Let $h_{ik} = h_{i(k-1)} - m_{i(k-1)}$.
   (f) If h_{ik} is a IMF then set $\mathrm{IMF_i} = h_{ik}$, else go to step (b) with $k = k + 1$.

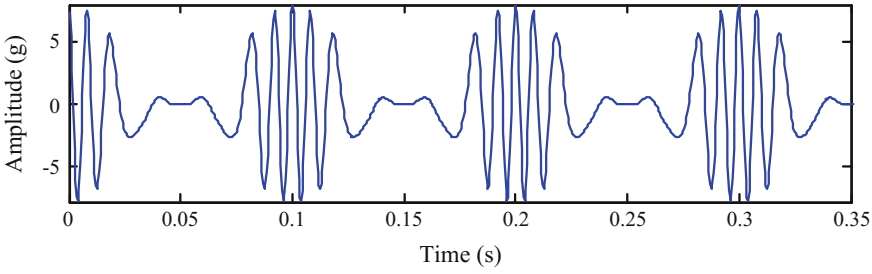


Fig. 1 Waveform of a typical IMF

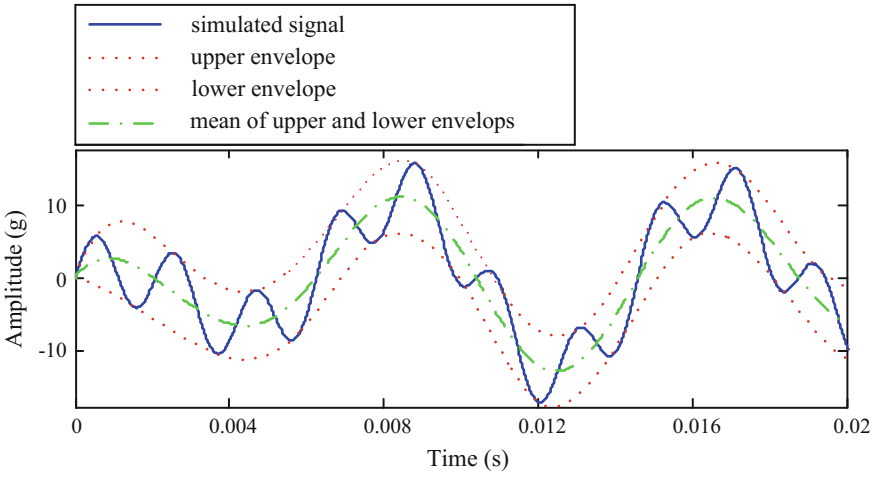


Fig. 2 Upper and lower envelops and their mean of a signal

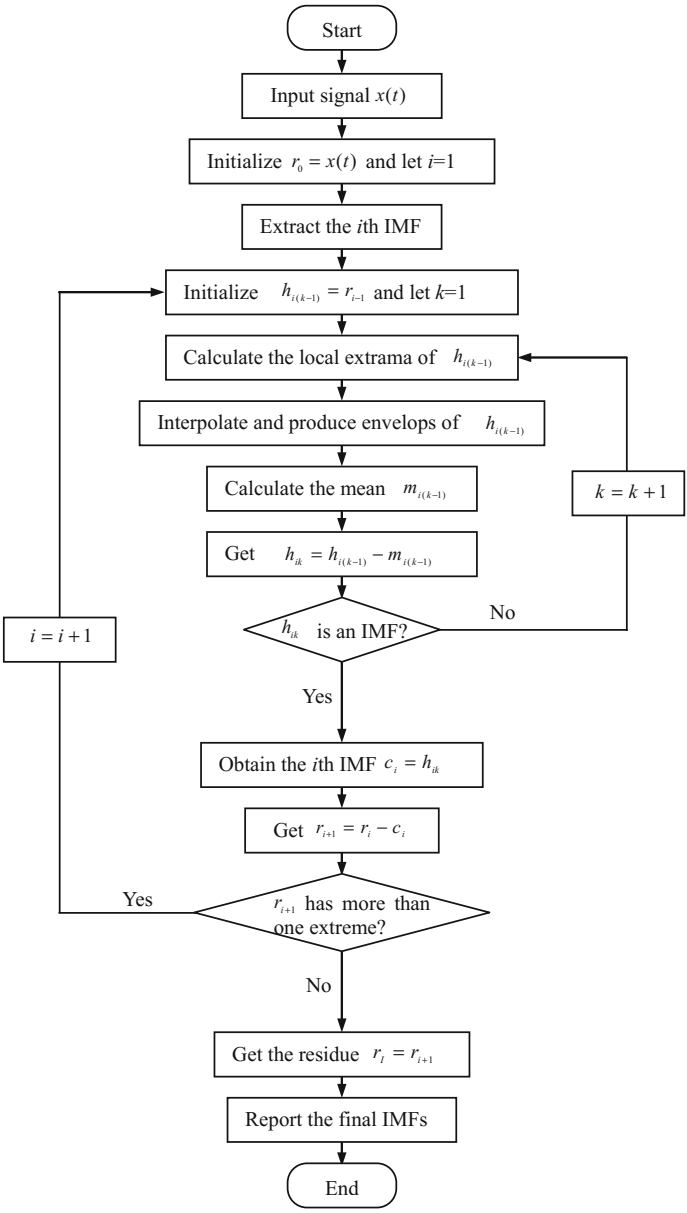


Fig. 3 Flow chart of EMD

(3) Define $r_{i+1} = r_i - \mathrm{IMF_i}$.
(4) If r_{i+1} still has least 2 extrema then go to step (2) else decomposition process is finished and r_{i+1} is the residue of the signal.

Thus, we can decompose the signal into $\boldsymbol{I}$ IMFs and a residue r_I, which is the mean trend of $x(t)$. Summing up all IMFs and the final residue r_I, we get $x(t) = \sum_{i=1}^{I} c_i + r_I$. The frequency bands of IMFs $c_1, c_2, \ldots, c_I$ ranges from high to low. The frequency components contained in each frequency band are different and they change with the variation of signal $x(t)$. Figure 3 shows the steps of the EMD algorithm.

A simulation is presented here to illustrate the decomposition results of EMD method. Given a signal $x(t)$, it consists of three components: a high-frequency sinusoidal wave, a low-frequency sinusoidal wave and a trend component. We use the EMD method to decompose this signal following the steps in Fig. 3. The decomposed components and the simulated signal $x(t)$ are shown in Fig. 4. From Fig. 4, it can be seen that two IMFs c_1 and c_2, and a residue r_2 are produced. Among them, c_1 and c_2 correspond to the two sinusoidal waves with different frequencies and the residue r_2 reflects the trend component embedded in the simulated signal.

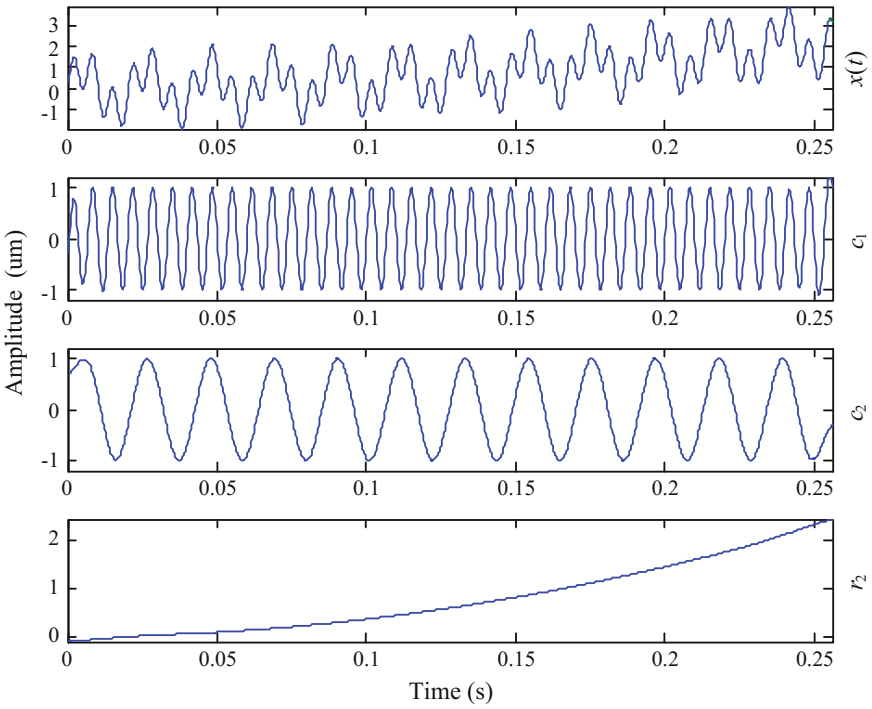


Fig. 4 Illustration of the EMD method

2.2 Problems of EMD

Although EMD largely contributes to the analysis of non-stationary and nonlinear signals, it also has weaknesses as well. For example, EMD produces end effects; the IMFs are not strictly orthogonal to each other; mode mixing sometimes occurs between IMFs.

(a) End effects

For a clear understanding of the end effects of EMD, we use the simulated signal shown in Fig. 4 to illustrate the end effects. We display the two sinusoidal waves included in the simulated signal and the decomposed IMFs of the simulated signal by EMD in Fig. 5. It is seen that there are distortions at the two ends of IMFs. This phenomenon is called end effects and it is caused by the EMD algorithm itself.

(b) Problem of orthogonality

The decomposed IMFs by EMD are not strictly orthogonal to each other. As we all know, if two components are orthogonal to each other, the dot product between them is zero. Here we also take the IMFs in Fig. 4 as an example. Calculating the dot product between the two IMFs c_1 and c_2, we obtain the value of 1.5 instead of zero. This means that IMFs c_1 and c_2 are not strictly orthogonal to each other. Moreover, the energy of the two IMFs and the residue can be calculated as 514.6,

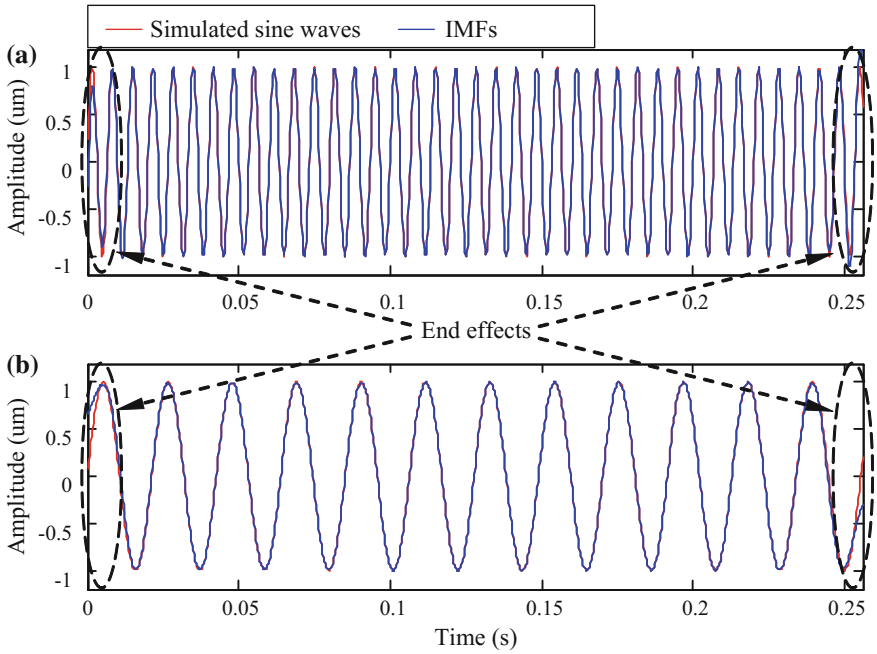


Fig. 5 **a** Simulated high-frequency sine wave and IMF c_1, and **b** simulated low-frequency sine wave and IMF c_2

515.2 and 1161.8, respectively, which means that the total energy of the three decomposed components is 2191.6. It is not equal to the energy of the simulated signal 2125.8. This indicates that when a signal is decomposed by EMD, the energy is not conservative before and after decomposition.

(c) Mode mixing

EMD method has another obvious shortcoming called mode mixing. The mode mixing of EMD is defined as a single IMF including oscillations of dramatically disparate scales, or a component of a similar scale residing in different IMFs.

To illustrate the problem of mode mixing in EMD, another simulated signal $x(t)$ is considered in this section. The simulated signal is shown in Fig. 6a. There is a sine wave of 36 Hz and small impulses included in this simulated signal. Therefore, it is a combined signal and actually consists of two components. Utilizing EMD on the signal, the decomposed results are shown in Fig. 6.

From Fig. 6, we can see that mode mixing is occurring between IMFs c_1 and c_2 since there are neither indications of a sinusoidal wave nor indications of small impulses. The sinusoidal wave and the impulses are decomposed into the same IMF (c_1). That is to say, these two IMFs obtained by EMD are distorted obviously and both IMFs c_1 and c_2 of EMD fail to represent the characteristics of signal $x(t)$ accurately. This is a typical problem of mode mixing.

Mode mixing of EMD is a result of signal intermittency. To solve the problem of mode mixing in the original EMD, ensemble empirical mode decomposition (EEMD), was developed by Wu and Huang by adding noise to the investigated signal [20]. A brief introduction of EEMD is given in the next section.

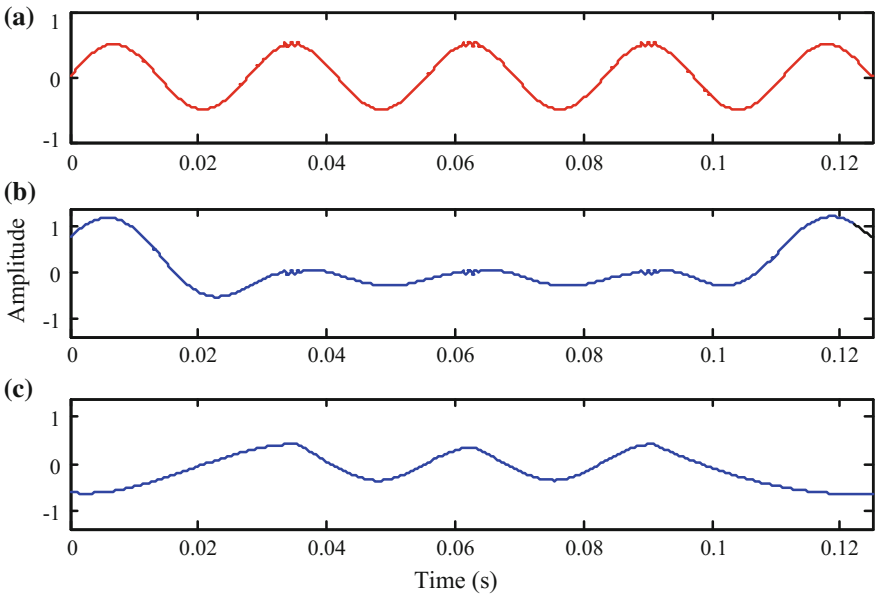


Fig. 6 Decomposition result with EMD: **a** the simulation signal, **b** IMF c_1, and **c** IMF c_2

Besides the end effects and mode mixing mentioned above, the EMD method has some other weaknesses, such as lacking a theoretical foundation, sifting stop criterion, extremum interpolation, etc. More details can be found in Refs. [14–16].

2.3 Hilbert-Huang Transform

Hilbert-Huang transform (HHT) mainly consists of two steps: EMD and Hilbert transform. EMD can decompose a signal into a collection of IMFs, which are almost monocomponent. Hilbert transform is defined as the convolution of signal $x(t)$ with $1/t$, shown in Eq. (1). Through the Hilbert transform, local properties of $x(t)$ are emphasized.

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

Combining $x(t)$ and $y(t)$,we can obtain the analytic signal $z(t)$ of $x(t)$

$$\begin{cases} z(t) = x(t) + iy(t) = a(t)e^{i\phi(t)} \\ a(t) = \sqrt{x^2(t) + y^2(t)} \\ \varphi(t) = \arctan(y(t)/x(t)) \end{cases} \tag{2}$$

where $a(t)$ is the instantaneous amplitude of $x(t)$, which reflects how the energy of $x(t)$ varies with time t, and $\varphi(t)$ is the instantaneous phase of $x(t)$. If the signal $x(t)$ is monocomponent, then the time derivative of instantaneous phase $\varphi(t)$ will be the physical meaning of instantaneous frequency $\omega(t)$ of the signal $x(t)$. Then the instantaneous frequency $\omega(t)$ is given as

$$\omega(t) = \frac{d\varphi(t)}{dt} \tag{3}$$

As discussed before, EMD can generate almost monocomponent IMFs, which provides an opportunity for the instantaneous frequency applied to complicated signals. For the signal $x(t)$, I IMFs are produced by EMD. Applying the Hilbert transform to each IMF, and calculating the instantaneous frequency and amplitude, we can express signal $x(t)$ in the following representation:

$$x(t) = \sum_{i=1}^{I} a_i(t) \exp\left(j \int \omega_i(t) dt\right) \tag{4}$$

Therefore, based on the IMFs obtained by EMD, the Hilbert transform generates a time-frequency-energy distribution to depict signal $x(t)$. The EMD-based Hilbert transform is called Hilbert-Huang transform (HHT).

3 Improved EMD Methods

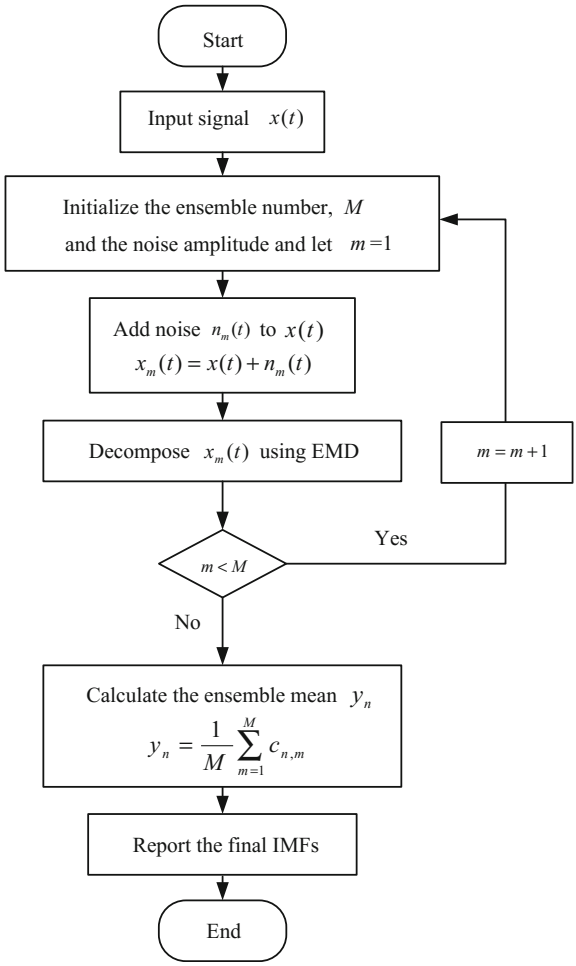
3.1 EEMD Method

As we all know, mode mixing is a typical problem of EMD method. When the problem of mode mixing occurs, an IMF can cease to have physical meaning by itself, suggesting falsely that there may be different physical processes represented in a mode [3]. To overcome this problem, ensemble empirical mode decomposition (EEMD) was proposed based on the statistical properties of white noise. In this new method, the true IMFs are defined as the mean of an ensemble of trials which consist of the decomposition results of the signal plus a normally distributed white noise with a constant standard deviation [20]. When the white noise is decomposed by EMD, EMD behaves likes a dyadic filter bank: the Fourier spectra of various IMFs collapse to a single shape along the axis of the logarithm of frequency [27]. In addition, the result provided by Flandrin, Gonçalvès, and Rilling demonstrated that the white noise could help data analysis in the EMD method [28]. In the process of EEMD, different white noise with zero mean and a constant standard deviation is added to the original signal and the combined signal is decomposed using EMD method in each trial. When the white noise is added to the signal, the components in different in different scales of the signal are automatically projected onto proper scales of reference established by the white noise in the background. The influence of the added noise can be decreased or even completely canceled out in the ensemble mean of enough trials. Therefore, the ensemble mean is treated as the true answer for the reason that only the signal is reserved when more and more trials are carried out in the ensemble process. The principle of EEMD advanced here is on the basis of the observations in the following [20].

(1) A collection of white noise cancels each other out in an ensemble mean; hence, only the signal can be reserved in the final noise-added signal ensemble mean.
(2) White noise is used to force the ensemble to find all possible solutions; it makes the signals of different scale reside in the corresponding IMFs, and the resulting ensemble mean can be more meaningful.
(3) The decomposition with truly physical meaning of EMD is not the one without noise; it is designated to be the ensemble mean of a large number of trials consisting of the noise-added signal. More detailed description of EEMD can be found in Ref. [20].

Based on the principle and observations as mentioned earlier, the EEMD algorithm is given below and Fig. 7 is its flow chart.

Fig. 7 Flow chart of EEMD



1. Initialize the number of ensemble, ***M***, the amplitude of the added white noise, and $m = 1$.
2. Perform the m-th trial on the signal added white noise.
   (a) Add a white noise series with the given amplitude to the signal to be studied
   $x_m(t) = x(t) + n_m(t)$, where $n_m(t)$ indicates the m-th added white noise series, and $x_m(t)$ represents the noise-added signal of the m-th trial.
   (b) Decompose the noise-added signal $x_m(t)$ into ***N*** IMFs, $c_{n,m}(n = 1, 2, \ldots, N)$, using EMD, where $c_{n,m}$ denotes the nth IMF of the m-th trial, and ***N*** is the number of IMFs.

(c) If $m < M$ then go to step (a) with $m = m + 1$. Repeat steps (a) and (b) again and again with different white noise series but having the same amplitude each time.

3. Calculate the ensemble mean y_i of the $\boldsymbol{M}$ trials for each IMF

$$y_n = \frac{1}{M}\sum_{m=1}^{M} c_{n,m}, \quad n = 1, 2, \ldots, N, \quad m = 1, 2, \ldots, M \tag{5}$$

(4) Report the mean y_n $(n = 1, 2, \ldots, N)$ of each of the $\boldsymbol{N}$ IMFs as the final IMF.

In order to demonstrate the improvement of EEMD method, the simulated signal in Fig. 6a is decomposed again using EEMD with the ensemble number 100 and the added noise amplitude 0.01 time standard deviation of the signal. The original signal is a sine wave of 36 Hz attached by small impulses. The decomposition results of EEMD method are shown in Fig. 8.

It has been concluded from Fig. 6 that the mode mixing is serious between the two IMFs obtained by EMD. However, it can be seen from Fig. 8b, c that the two components contained in the signal are decomposed into two IMFs perfectly using EEMD. IMF c_1 in Fig. 8b denotes the impulse components and IMF c_2 in Fig. 8c indicates the sine wave. Therefore, EEMD is able to overcome the mode mixing

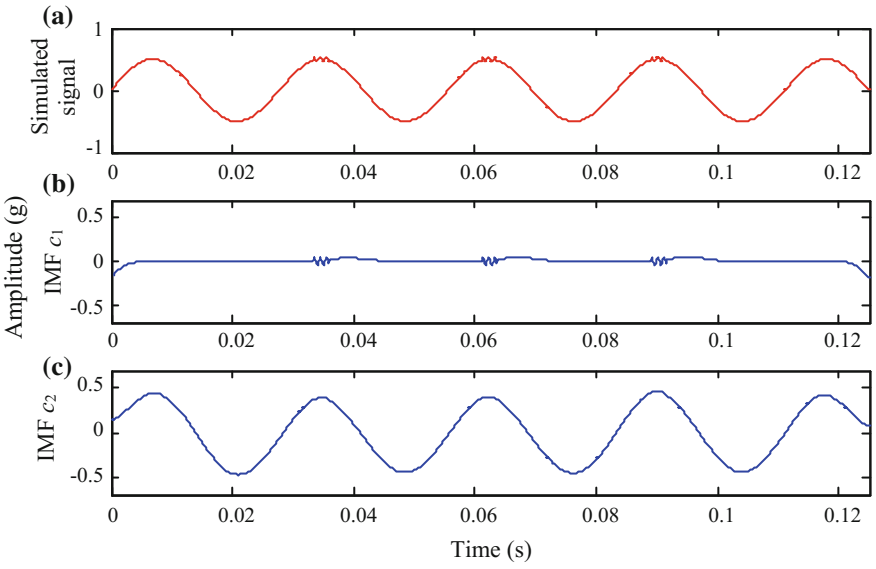


Fig. 8 **a** The simulated signal; **b** and **c** the decomposition results of EEMD

problem existing in EMD method and achieve an improved decomposition with physical meaning.

When EEMD method is used, parameters such as the number of ensemble and amplitude of the added noise need to be set reasonably. The following part will discuss the choice of the two parameters.

(a) The number of ensemble

The relationship between the ensemble number and the amplitude of the added white noise is given in the following equations [20].

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

or

$$\ln e + \frac{a}{2} \ln N = 0 \tag{7}$$

where N is the number of ensemble, a is the amplitude of the added white noise, and e is the standard deviation of error, which is defined as the difference between the input signal and the corresponding IMFs.

In the process of EEMD method, small amplitude of the added white noise may lead to a small error. However, if the amplitude of added noise is too small, it may not change the distribution of extrema that the EEMD method relies on. This is true when the investigated signal has a large gradient. Thus, the amplitude of the added noise should not be too small for the effectiveness of EEMD method. On the other hand, the error caused by the added white noise could always be reduced to a quite small even negligible level by increasing the number of ensemble. Generally, an ensemble number of a few hundred will lead to an exact result, and the remaining noise would cause less than a fraction of one percent of error if the added noise has the amplitude that is a fraction of the standard deviation of the investigated signal [20].

(b) The amplitude of the added white noise

The investigation in references indicated that EMD is a noise-friendly method [20]. In addition, increasing noise amplitudes and ensemble numbers changes the decomposition results little as long as the amplitude of added noise is moderate and the ensemble number is large enough.

The fact is that when the amplitude of noise increases, the number of ensemble should increase to reduce the influence of the added noise in the decomposed results. It is suggested that the amplitude of the added white noise is about 0.2 time standard deviation of the investigated signal [20]. However, it is not always the proper amplitude of the added noise for any cases. Generally, when the signal is dominated by high frequency components, the noise amplitude needs to be smaller. On the contrary, when the signal is dominated by low frequency

components, the noise amplitude should be larger. However, there is no a specific equation reported in the literature to guide the choice of the noise amplitude until now. Thus, for an investigated signal, different noise levels should be tried to select the appropriate one.

3.2 AEEMD Method

As stated above, EEMD, as a noise-assisted data analysis method, is aimed to solve the problem of mode mixing in EMD [20]. With the help of the added finite white noise, EEMD is supposed to eliminate the mode mixing problem [21]. The performance of EEMD, however, depends on the parameters adopted in the process of decomposition, such as the sifting number, and the amplitude of the added noise. In fact, these parameters were set as constant values whether the signal to be investigated contains high or low components in most current studies on EEMD [14]. Therefore, the problem of mode mixing is not solved completely and further work need to be done to improve the performance of EEMD.

On the basis of the investigation of the filtering behavior of EMD/EEMD and the relation between the signal frequency components and the amplitude of the added noise, a new adaptive ensemble empirical mode decomposition method (AEEMD) is proposed [24]. The new method adaptively selects the sifting number and decides the amplitude of the added noise according to the signal frequency components in decomposition process. By adopting the two parameters, the performance of EEMD is going to be improved in feature extraction and fault diagnosis.

In the process of EEMD, high and low frequency components have different sensitivity to noise. Therefore, larger noise and more sifting number had better be adopted when high-frequency IMFs are extracted, while smaller noise and less sifting number had better be used when low-frequency IMFs are extracted. To satisfy this requirement for noise, different kinds of noise are tried and tested. The result shows that the noise whose amplitude changes with its frequency in sine form performs best. Therefore, the noise of this form is constructed and utilized in

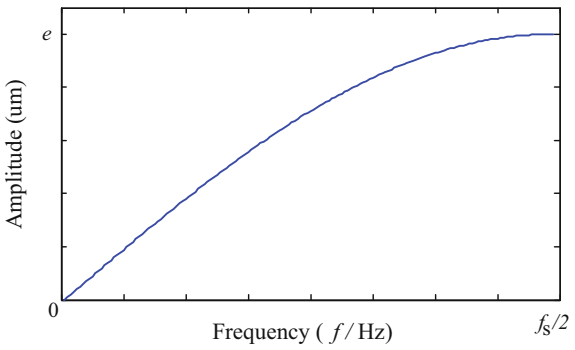


Fig. 9 Spectrum of the noise constructed

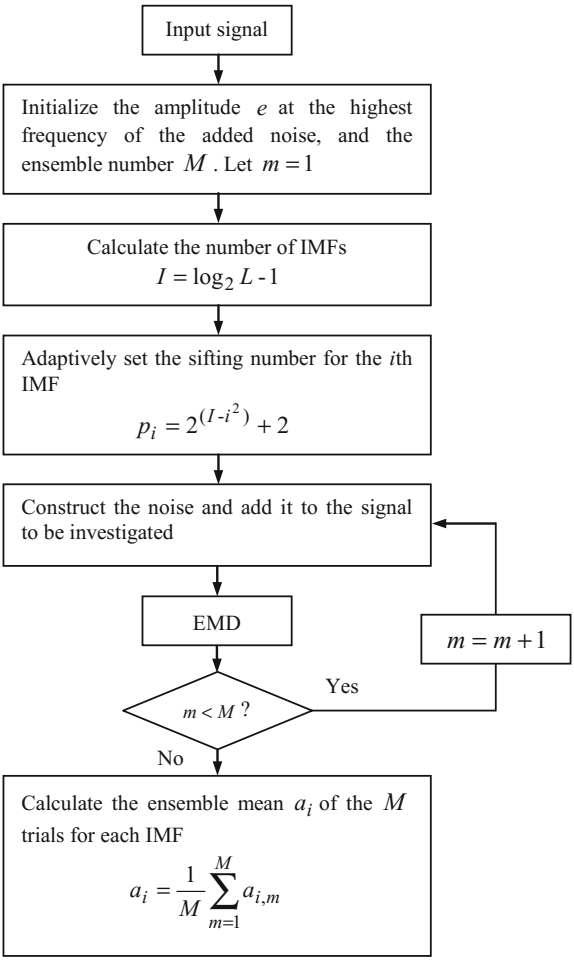


Fig. 10 Flowchart of AEEMD

AEEMD in place of white noise adopted in the original EEMD. Figure 9 gives the frequency spectrum of the constructed noise, in which f_s represents the sampling frequency and e denotes the amplitude at the highest frequency. On the other hand, EMD method is an effective self-adaptive dyadic filter bank when applied to white noise. Therefore, the sifting number for each IMF is adaptively set following Eq. (8). Figure 10 gives the flow chart of AEEMD algorithm. The concrete steps are as follows.

(1) Initialize the amplitude e of the highest frequency of the added noise, the number of ensemble M, generally $M = 100$ and $e = 0.2$. Let $m = 1$.
(2) Calculate the number of IMFs based on the signal length [20]

$$I = \log_2 L - 1 \tag{8}$$

where L is the signal length.

(3) Adaptively set the sifting number p_i for the i-th IMF according to the following equation.

$$p_i = 2^{(I-i^2)} + 2, \quad i = 1, 2, \ldots, I \tag{9}$$

(4) Construct the noise as shown in Fig. 9 and add it to the signal to be investigated.
(5) Perform EMD on the added-noise signal and obtain the m-th decomposition result $a_{i,m}$.
(6) If $m < M$ then go to step (4) with $m = m + 1$. Repeat steps (4) and (5).
(7) Calculate the ensemble mean a_i of the M trials for each IMF and report the mean as the final IMF.

$$a_i = \frac{1}{M}\sum_{m=1}^{M} a_{i,m}, \quad i = 1, 2, \ldots, I, \quad m = 1, 2, \ldots, M \tag{10}$$

To demonstrate the effectiveness of AEEMD method, a simulation signal is constructed. For the reason that modulation and impact are two typical fault events in rotating machinery, the simulation signal contains modulation as well as impact components. What is more, it also consists of a high-frequency sinusoidal wave and a low-frequency sinusoidal wave respectively to represent certain rotating frequencies of machinery. Therefore, there are four components having different physical meaning in the simulation signal. The four components and the simulation signal combined by them are shown in Fig. 11a–e, respectively.

AEEMD method is used to decompose the simulation signal and the decomposition results are shown in Fig. 12. It can be seen from the result that IMFs 1–4 correspond to the impact component, the modulation component, the high-frequency sinusoidal wave and the low-frequency sinusoidal wave respectively.

Comparing the decomposed IMFs in Fig. 12 with the real components in Fig. 11a–d, it can be inferred that the different components embedded in the simulation signal are extracted accurately by AEEMD. For comparison, the simulation signal is analyzed using the original EMD too and the decomposition result is displayed in Fig. 13. It is seen that the problem of mode mixing between different components is very serious and there are distortions for some IMFs. For example, the first IMF contains not only the impact component but also the modulation component. This result illustrates that the original EMD fails to produce the reasonable decomposition. Based on the above simulation and comparison, it could be inferred that AEEMD performs more effective than the original EMD, by adding noise with the amplitude varying as a sinusoidal relation with its frequency into the signal, and adaptively changing the sifting number for different IMFs.

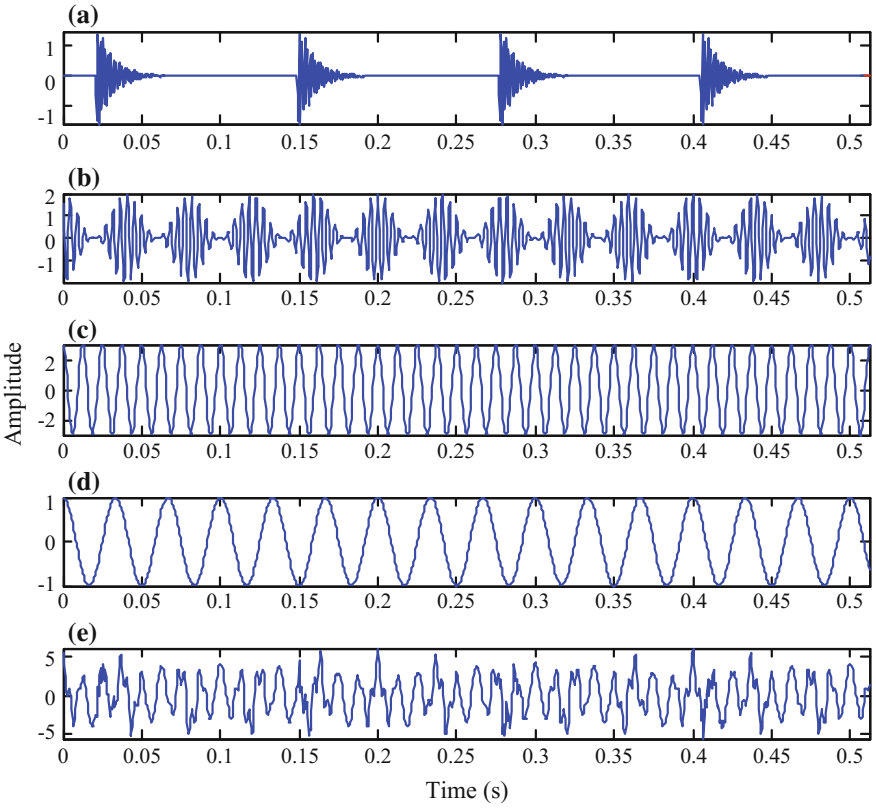


Fig. 11 **a**–**d** Four components, **e** Simulation signal

3.3 CEEMDAN Method

EEMD is mainly to alleviate the problem of mode mixing caused by EMD [20, 29], however, it still has some shortcomings. For example, the EEMD method decomposes a signal adding the white Gaussian noise, and the final IMFs are obtained by averaging the IMFs. This would probably lead to some residual noise in the reconstructed signal. In addition, if the white Gaussian noise in each decomposition process is added with different amplitudes, it probably may produce a different number of IMFs, which makes it difficult for the averaging [30, 31].

To overcome the above shortcomings of EEMD, Torres et al. proposed an algorithm called a complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) [30]. Furthermore, Colominas et al. continued to make improvements on CEEMDAN [31]. In this improved CEEMDAN method, a particular noise $E_k(w^{(i)})$ instead of the white Gaussian noise is added at each stage of the decomposition, where $E_k(w^{(i)})$ means the kth IMF of the white Gaussian noise

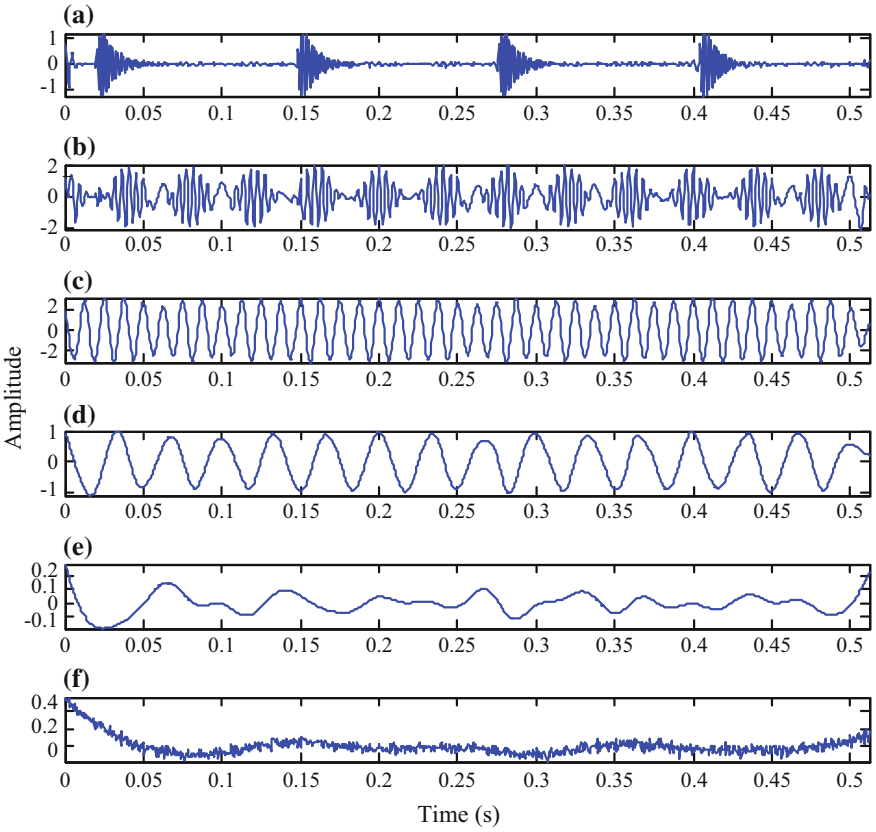


Fig. 12 IMFs using AEEMD for the simulation signal

decomposed by EMD. Moreover, this method defines the true IMF as the difference between the current residue and the average of its local means. As a result, the problem of remaining noise in IMFs is much alleviated and the problem of the final averaging because of a different number of IMFs is solved.

Let $M(\cdot)$ represent the operator that produces the local means of the signal x, and $E_k(\cdot)$ be the operator which produces the kth IMF decomposed by EMD. Obviously, there exists a relation that $E_1(x) = x - M(x)$. Considering the relation between the first IMF c_1 and the residue r_1:$c_1 = x - r_1$, $c_1 = \frac{1}{I}\sum_{i=1}^{I} E_1(x) = x - \frac{1}{I}\sum_{i=1}^{I} M(x^{(i)})$, where $\boldsymbol{I}$ means the averaging number of IMFs, there exists $\frac{1}{I}\sum_{i=1}^{I} M(x^{(i)}) = r_1$. The decomposition using CEEMDAN is based on the following principles [30, 31] and a flow chart of the CEEMDAN algorithm is shown in Fig. 14.

Step 1. Add $E_1(w^{(i)})$ to the original signal x, $x^{(i)} = x + \beta_0 E_1(w^{(i)})$, where $w^{(i)}$ indicates the ith added white noise.

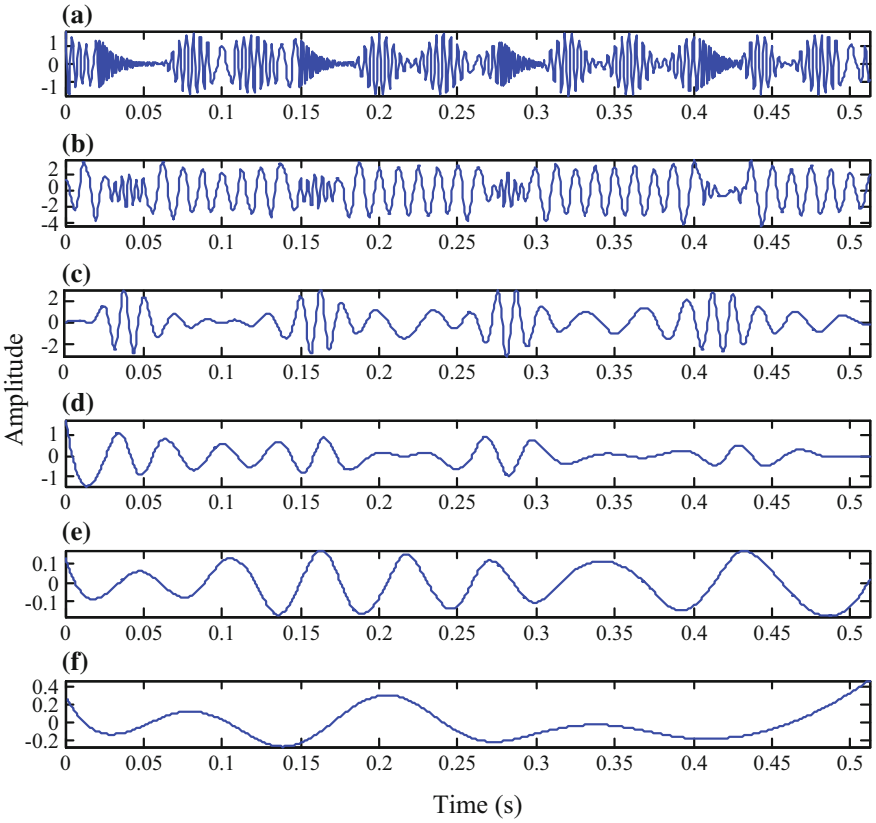


Fig. 13 IMFs using the original EMD for the simulation signal

Step 2. Use EMD to calculate the local means of $x^{(i)}$ and average them for the first residue, $r_1 = \frac{1}{I}\sum_{i=1}^{I} M(x^{(i)})$, then calculate the first IMF c_1 as $c_1 = x - r_1$.

Step 3. Obtain the second IMF c_2 as $c_2 = r_1 - r_2$, where $r_2 = \frac{1}{I}\sum_{i=1}^{I} M(r_1 + \beta_1 E_2(w^{(i)}))$.

Step 4. Similarly, the k-th IMF c_k is computed as $c_k = r_{k-1} - r_k$, where $r_k = \frac{1}{I}\sum_{i=1}^{I} M(r_{k-1} + \beta_{k-1} E_k(w^{(i)}))$, $k = 2, 3 \ldots \mathrm{N}$.

The coefficients β_k represent the selection of the SNR at each stage, where $\beta_k = \varepsilon_0 \mathrm{std}(r_k)$.

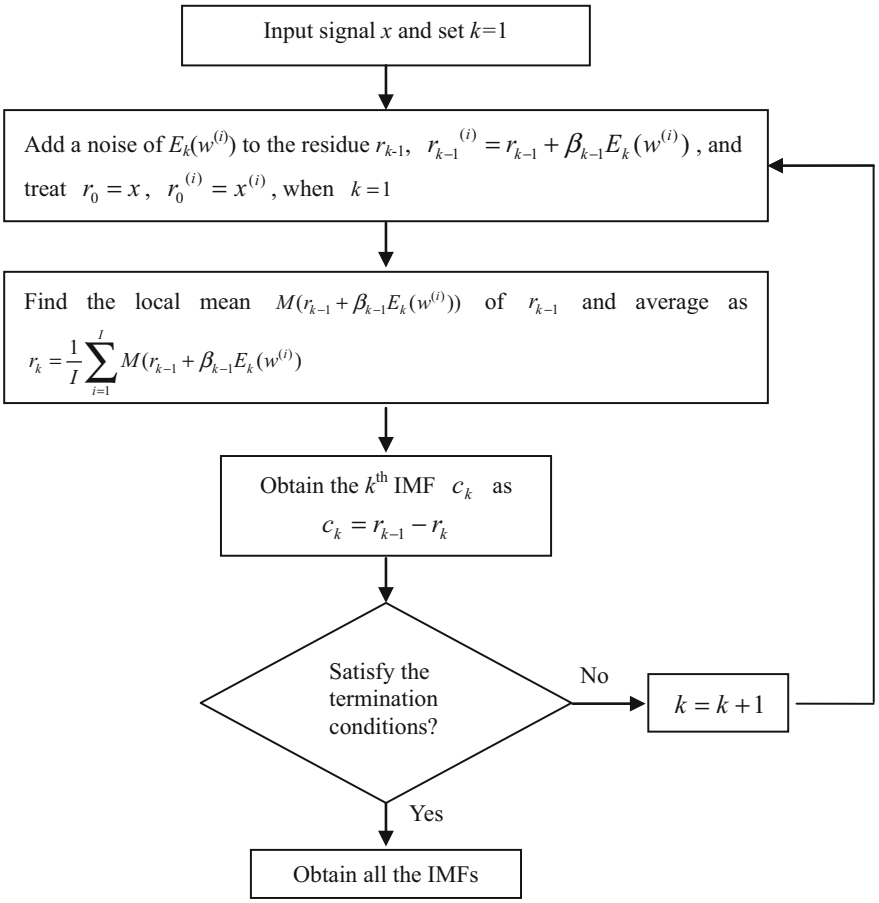


Fig. 14 Flow chart of the CEEMDAN algorithm

To illustrate the decomposition difference between EEMD and CEEMDAN, a simulated signal $x(t)$ is implemented here. There are four components involved in this signal: a high-frequency sinusoidal wave, a low-frequency sinusoidal wave, an impact component and a modulation component. The simulated signal and the four components are shown in Fig. 15a–e, respectively.

According to Refs. [20, 32], when the value of ε_0 is close to 0.2, it often has a remarkable performance of the decomposition results. Consequently, we choose the noise amplitude $\varepsilon_0 = 0.2$ and the ensemble size $I = 100$ in the decomposition of CEEMDAN. The decomposed results of the simulated signal using EMD method and the CEEMDAN method are shown in Figs. 16 and 17, respectively. The components (a–d) in Fig. 16 correspond to high-frequency sinusoidal wave, low-frequency sinusoidal wave, the impact and the modulation component, respectively. It can be seen that the high-frequency sinusoidal wave is mixed with

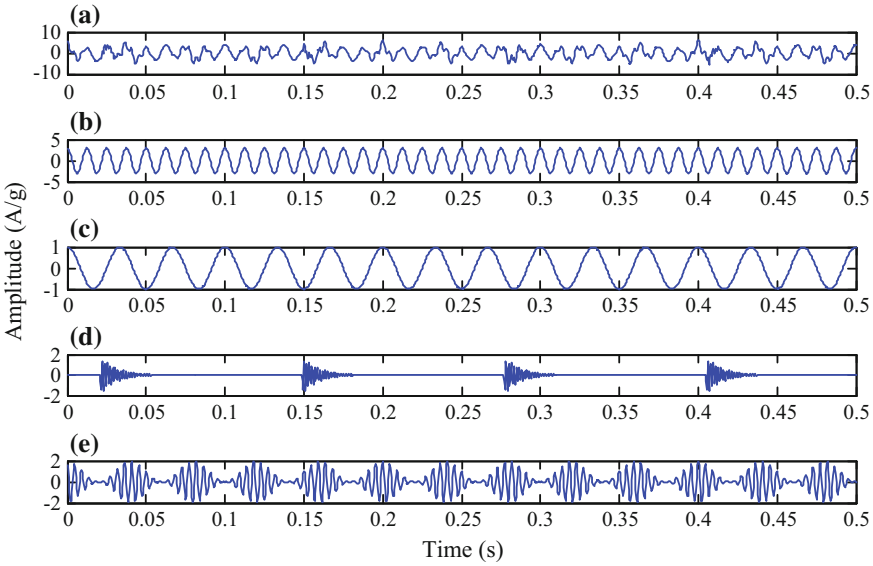


Fig. 15 Four components and the simulated signal: **a** The simulated signal and **b–e** The four components

the low-frequency sinusoidal wave and the impact component is mixed with the modulation component obviously. In Fig. 17, it is seen that the individual components hidden in the simulated signal can be extracted using the method based on CEEMDAN. Especially, the impact component and the modulation component are presented clearly in the third and the forth IMF with an accurate waveform, respectively.

4 Fault Diagnosis of Rotating Machinery Using EMD Based Methods

4.1 Fault Diagnosis of Rotors

A power generator plays an important role in energy supply. It has a great meaning to diagnose the faults occurring in the power generator to guarantee the regular energy supply, avoiding the economic loss and saving the production cost. A structure sketch of a power generator in a thermal-electric plant in China is given in Fig. 18. This machine set is composed of a high pressure cylinder, a low pressure cylinder, a motor and an exciter.

A certain day, it was found that the high pressure cylinder vibrated so intensely that the virtual value of vibration signal exceeded the safety threshold, and then the

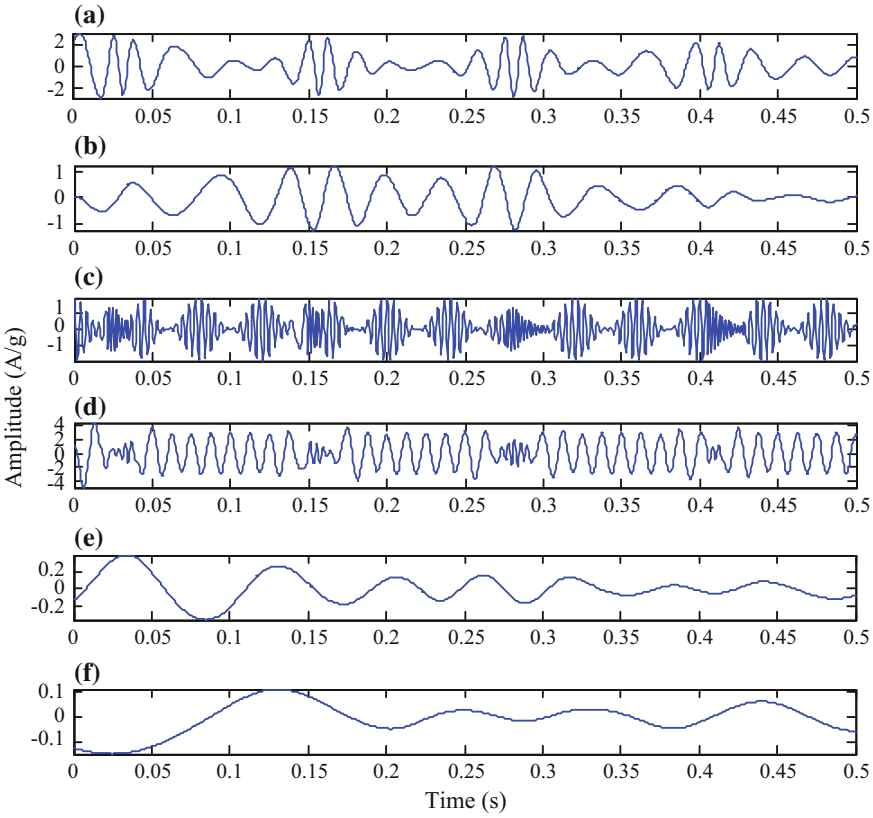


Fig. 16 Decomposed components of the simulated signal using EMD

online monitoring system began to sound the alarm. In the later one month, the machine vibrated even more violently. When the power generator was stopped to be maintained, it was found that one of the bearing bushes of the machine set had been broken. In order to identify the fault pattern, the vibration signal was collected by a vibration velocity transducer fixed on the high pressure cylinder, which is shown in Fig. 19. The signal length is 1024, and the sampling frequency is 2000 Hz. The rotating frequency of the machine set is 50.78 Hz.

First, the vibration signal was decomposed using EMD method, and the first six IMFs of the decomposed results are given in Fig. 20. A series of impulses could be seen in some local components of IMFs c1 and c2. Therefore, it can be inferred that periodic impacts occur in the high pressure cylinder.

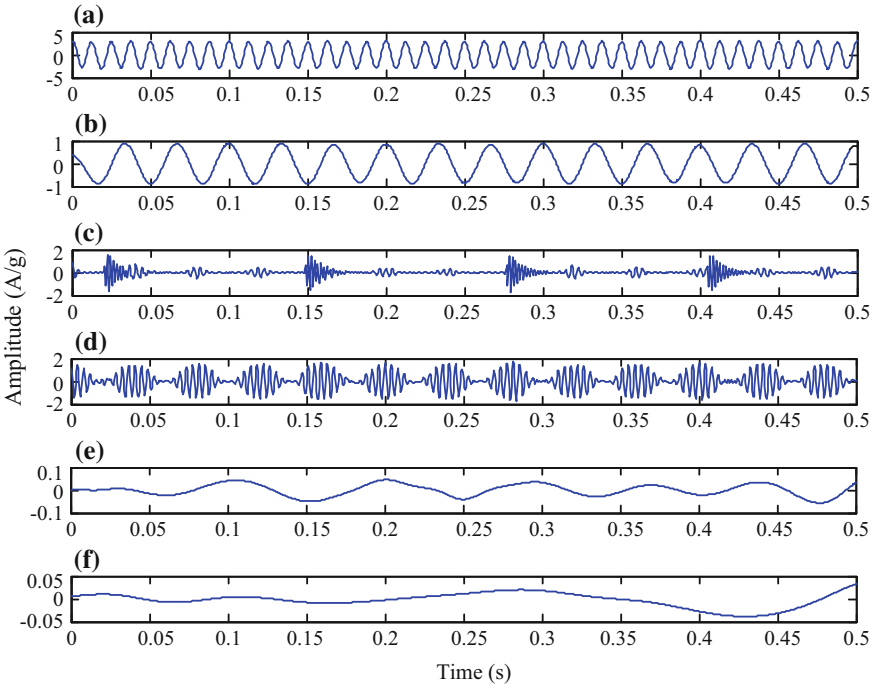


Fig. 17 Decomposed components of the simulated signal using the CEEMDAN method

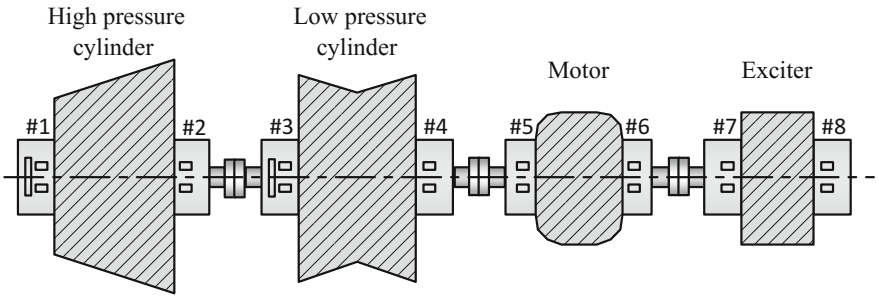


Fig. 18 Structure sketch of the power generator

However, what is the reason that caused the impacts between the rotor and the bearing bushes? Unfortunately, it is very difficult to answer this question according to the information provided by the IMFs of EMD owing to mode mixing occurring between different IMFs. In addition, there is no more fault information to clarify the fault cause.

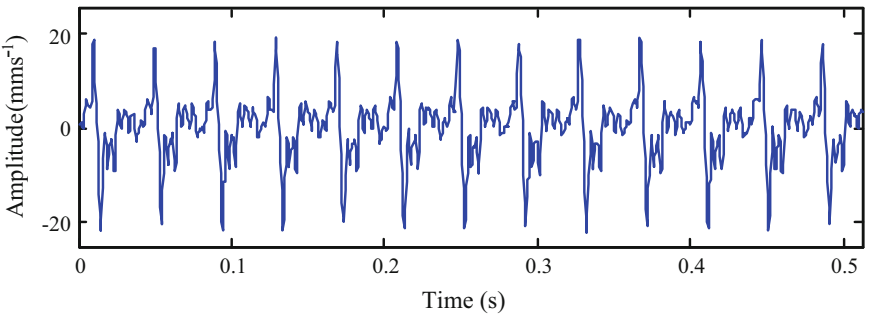


Fig. 19 Vibration signal collected from the high pressure cylinder

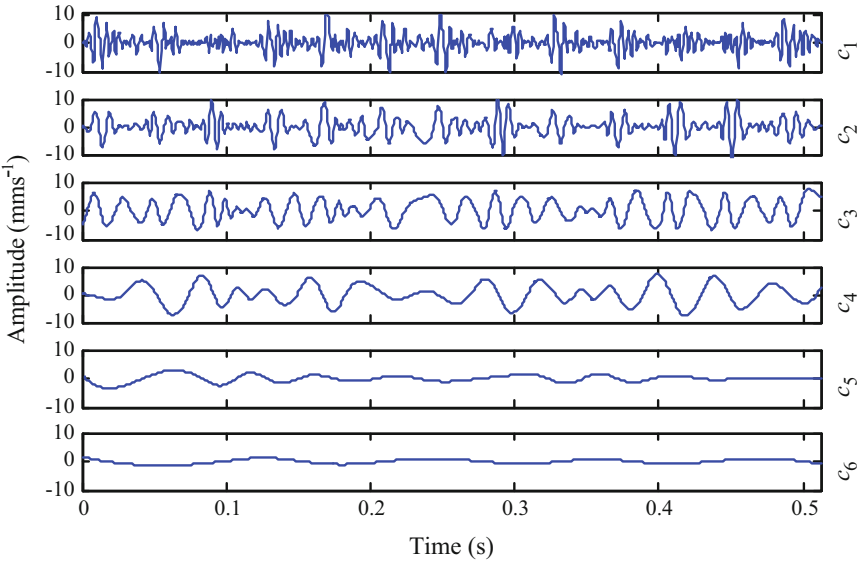


Fig. 20 Decomposed result of the vibration signal of the high pressure cylinder with EMD

To overcome the above difficulty, the CEEMDAN method with the ensemble number of 100 and the white noise amplitude of 0.3 time standard deviation of that of the vibration signal is applied to the signal decomposition. The decomposition results are given in Fig. 21. It can be clearly seen that each IMF has its real physical meaning. IMF c_1 corresponds to the added white noise. IMFs c_2 and c_3 indicate impulse components. IMF c_4 is the rotating frequency component of the machine

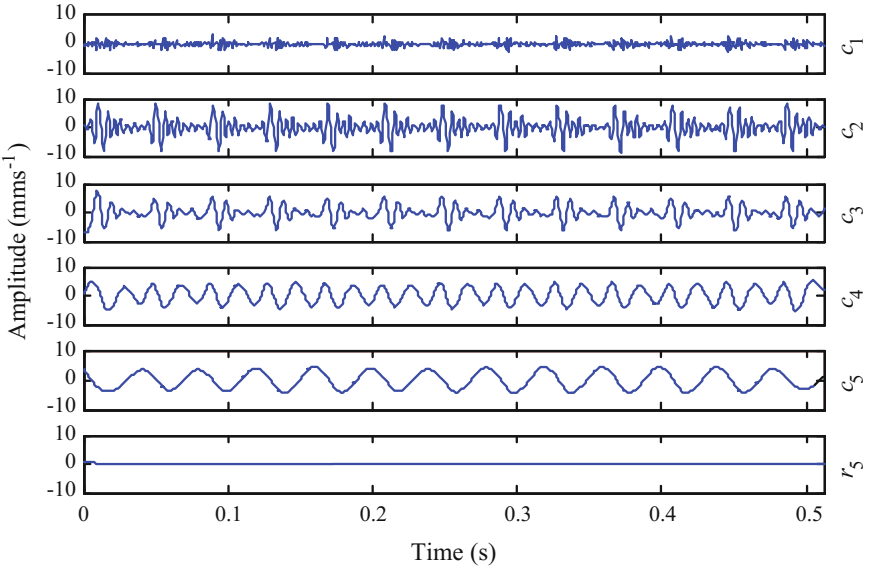


Fig. 21 Decomposed result of the vibration signal of the high pressure cylinder with EEMD

set whose value is 50.78 Hz. IMF c_5 is a component of 25.39 Hz, which is the half rotating frequency of the machine set.

It can be inferred that the fault pattern is not oil whirl in this machine set because oil whirl usually manifests itself by frequencies ranging from 42 to 48% of the rotating frequency of the rotor. In a rotor system, both looseness and rub behave themselves by the half rotating frequency. For the reason that there are impacts between the rotor and the bearing bushes, we can conclude that the fault of the power generator is the rub-impact pattern. That implies that the rotor system of the high pressure cylinder rub and at the same time impact the bearing bushes when the power generator is operating. Then impulse components are generated. Finally, the intense impacts broke one of the bearing bushes.

4.2 Fault Diagnosis of Gears

In modern industry, planetary gear boxes are widely used as a kind of special gear transmission structures owing to their advantages such as large transmission ratio, strong load-bearing capacity. They have big difference with fixed-axis gearboxes

Table 1 Parameters and characteristic frequencies of the planetary gearbox

Tooth number of gears			Gear number	Rotating frequency/Hz			Mesh frequency/Hz
Sun	Planetary	Ring	Planetary	Sun	Planetary	Carrier	
20	40	100	3	20	8.33	3.33	333.33

and exhibit unique behaviors, which increase the difficulty of fault diagnosis in planetary gearboxes [33–35].

In this section, experiments are conducted on a planetary gearbox test rig and vibration signals are collected to demonstrate the effectiveness of the adaptive EEMD in diagnosing gear faults. The planetary gearbox test rig consisted of two gearboxes, a 3-hp motor for driving the gearboxes, and a magnetic brake for loading. There were a planetary gearbox and a fixed-axis gearbox in the test rig. An inner sun gear is surrounded by several rotating planet gears, and a stationary outer ring gear in the planetary gearbox, which is our concern [33]. To simulate gear faults, a crack at the tooth root of one planetary gear is created in our experiments.

An accelerometer is fixed on the planetary gearbox casing to collect the vibration signals. The motor speed is about 20 Hz and the sampling frequency is set as 5120 Hz. The experimental parameters and the characteristic frequencies of the planetary gearbox are shown in Table 1. It can be seen from the table that the rotating frequency of one planetary gear is 2.5 times as large as that of the carrier. Therefore, when the carrier rotates 2 cycles, the planetary gear meshes 5 periods

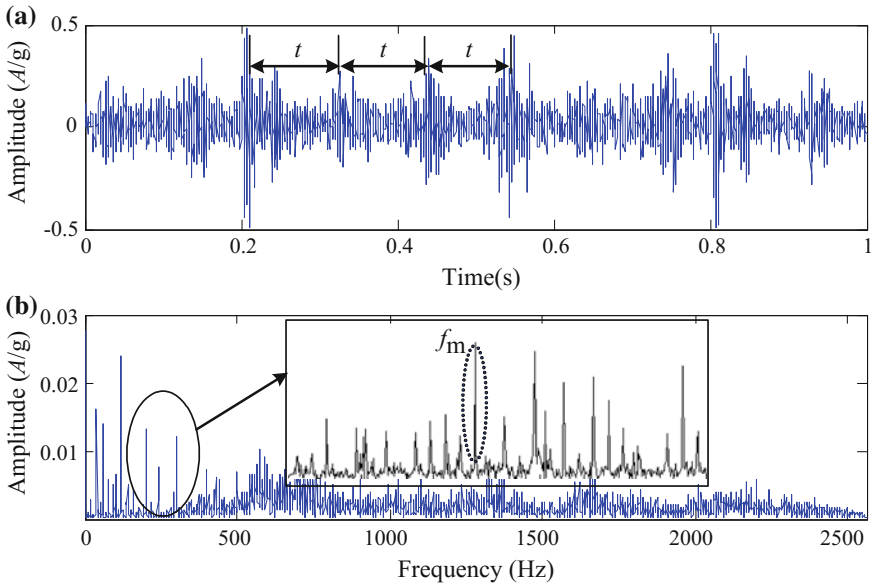


Fig. 22 Experimental signal **a** time-domain waveform, and **b** frequency spectrum

with the ring gear, i.e. 200 teeth. This tooth number is twice as large as that of the ring gear. That is to say, the ring gear meshes 2 periods with the planetary gear. In other words, the planetary gear returns to the initial position when the carrier rotates 2 cycles. For the carrier to finish rotating 2 cycles, it takes 2/3.33 = 0.6 s.

The vibration signal collected from the test rig with the cracked planetary gear is shown in Fig. 22a and its frequency spectrum is shown in Fig. 22b. It can be seen that there are a series of impulses in the time-domain waveform. The period of the impulses is nearly $t = 0.1$ s which means that the frequency of the impulses is about 10 Hz. There are three planetary gears in planetary gearbox and they pass the fixed accelerometer in turn. Therefore, the pass frequency of the planetary gears equals 3 times as large as the rotating frequency of the carrier, i.e. 10 Hz. It is apparent that the impulses in the time-domain waveform are caused by the rotation of the carrier, and they are normal vibration components of the gearbox. Unfortunately, it is difficult to extract any fault characteristics besides these impulses for the reason that the fault features of the planetary gearbox are buried by the normal vibration components. The frequency spectrum of the vibration signal in Fig. 22b shows that there are rich sidebands around the mesh frequency and the interval of the sidebands is 3.33 Hz, which equals the rotating frequency of the carrier. Obviously, it is not the fault characteristics either. Therefore, the fault features of the cracked planetary gear cannot be found from the time-domain waveform or its frequency spectrum.

The adaptive EEMD method is used to process the above signal to extract the fault features of the cracked planetary gear. Among the IMFs decomposed by adaptive EEMD, the first IMF contains the richest information and consequently it is selected for further analysis. Figure 23 shows the detail of IMF1 and it can be seen that there are impulses with the period $T = 0.6$ s. It can be concluded that once

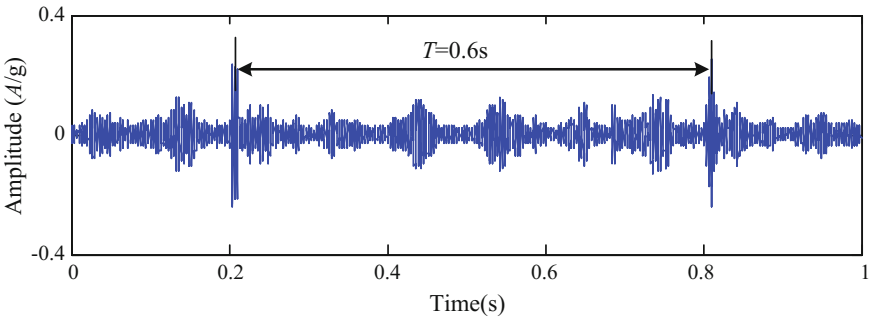


Fig. 23 First IMF extracted by the adaptive EEMD method

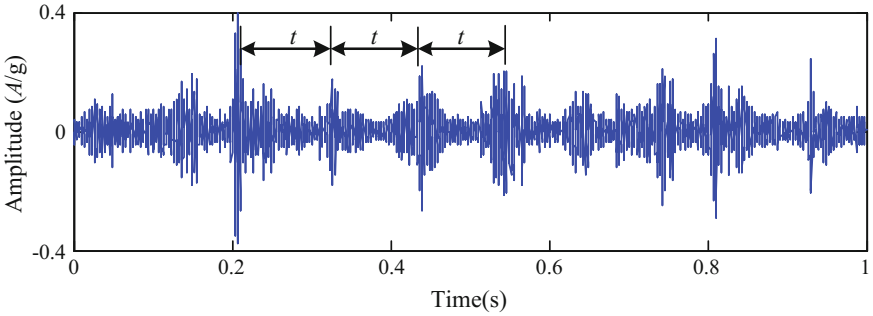


Fig. 24 First IMF extracted by the original EEMD method

the carrier rotates 2 cycles, the cracked planetary gear returns to the initial position. As a result, the fault period of the cracked planetary gear doubles the rotating period of the carrier, i.e. 0.6 s. That means that the impulse component with the period $T = 0.6$ s is resulted by the cracked planetary gear. Therefore, the adaptive EEMD method can extract the fault characteristics effectively. For comparison, the same signal is decomposed by original EEMD and the first IMF is shown in Fig. 24. In can be seen that there are also periodic impulses in the waveform of the IMF, the impulse ($T = 0.6$ s) caused by the cracked gear and those ($t = 0.1$ s) caused by the rotation of the carrier. However, these impulses are decomposed in the same IMF which means the mode mixing happens. It can be concluded from the comparisons that the adaptive EEMD is more effective than the original EEMD in fault characteristics extraction of the planetary gearbox.

4.3 Fault Diagnosis of Rolling Element Bearings

In this section, to demonstrate the effectiveness of the CEEMDAN based method, an experiment on a test bench of locomotive rolling element bearings was conducted. The detailed information of the test bench can be seen in Ref. [2]. The parameters and fault characteristic frequencies of the bearings are listed in Table 2.

It is known that even though a serious single fault occurs on the bearing, periodical impulses characterizing the fault can be submerged by the heavy noise, which may be even worse for compound faults. This is mainly because in rolling element bearings, once compound faults occur, different fault characteristics always couple with each other. Then the fault characteristics of compound faults turned to be complicated and difficult to be extracted, and the common used methods like

Fourier spectrum analysis probably will fail to work. Thus for the demonstration of the CEEMDAN method in the fault diagnosis of bearings, we choose a bearing with compound faults on the outer race and the roller.

A vibration signal was collected from the test bench with a sampling frequency of 12800 Hz and a signal length of 16,384 data points. The waveform of the signal and its corresponding Fourier spectrum is given in Fig. 25. The outer race and the roller fault characteristic frequencies of the bearing can be calculated as 44.5 and 19.94 Hz, respectively, since the rotation speed is 370 rpm. Obviously, in Fig. 25,

Table 2 Parameters and fault characteristic frequencies of the locomotive rolling element bearing

Bearing specs	52732QT
Inner race diameter D_i (mm)	160
Outer race diameter D_o (mm)	290
Roller diameter d (mm)	34
Roller number n	17
Contact angle α (deg)	0
Bearing rotating frequency (Hz)	f_r
Pitch diameter D (mm)	$D = \frac{1}{2}(D_\mathrm{i} + D_\mathrm{o})$
The inner race fault characteristic frequency f_inner	$f_\mathrm{inner} = \frac{1}{2} f_\mathrm{r}(1 + \frac{d}{D}\cos\alpha)n$
The outer race fault characteristic frequency f_outer	$f_\mathrm{outer} = \frac{1}{2} f_\mathrm{r}(1 - \frac{d}{D}\cos\alpha)n$
The roller fault characteristic frequency f_roller	$f_\mathrm{roller} = \frac{1}{2} f_\mathrm{r}[1 - (\frac{d}{D})^2 \cos^2\alpha]\frac{D}{d}$

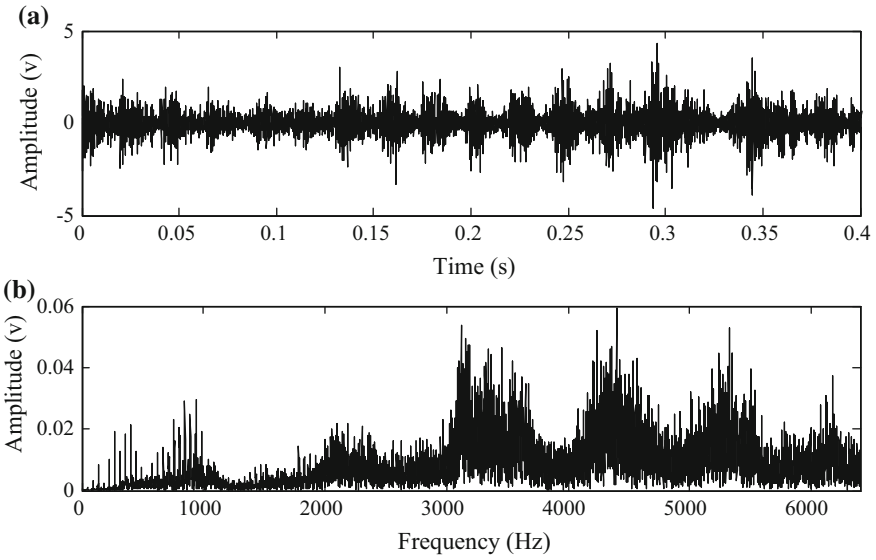


Fig. 25 **a** Vibration signal of a bearing with compound faults, and **b** Fourier spectrum

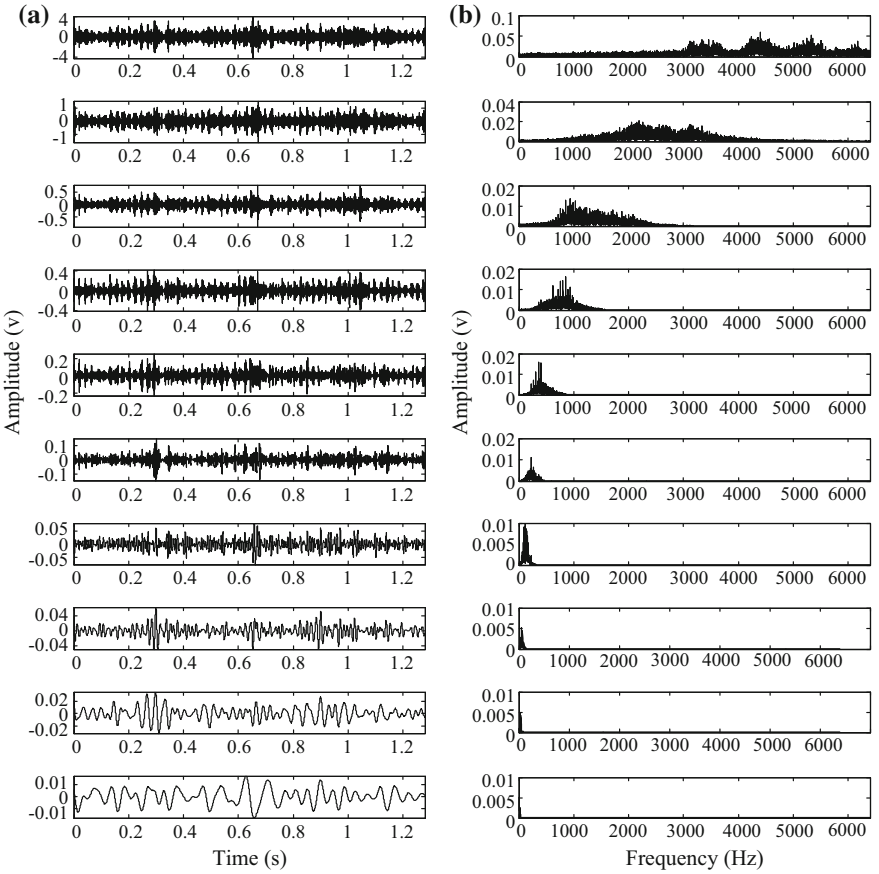


Fig. 26 Decomposition results of the vibration signal of a bearing with compound faults using EEMD: **a** The IMFs, and **b** The Fourier spectra of IMFs

neither the periodic impulses related to faults in the time-domain waveform nor the fault characteristic frequencies in the Fourier spectrum can be observed. The Fourier spectrum analysis fails to detect the compound faults.

To reveal the fault characteristics, EEMD is used to extract the fault characteristics from the signal. 16 IMFs are obtained after the decomposition. However, there is no obvious fault indication at the characteristic frequency of 44.5 and 19.94 Hz from the first 10 IMFs and their corresponding Fourier spectra in Fig. 26.

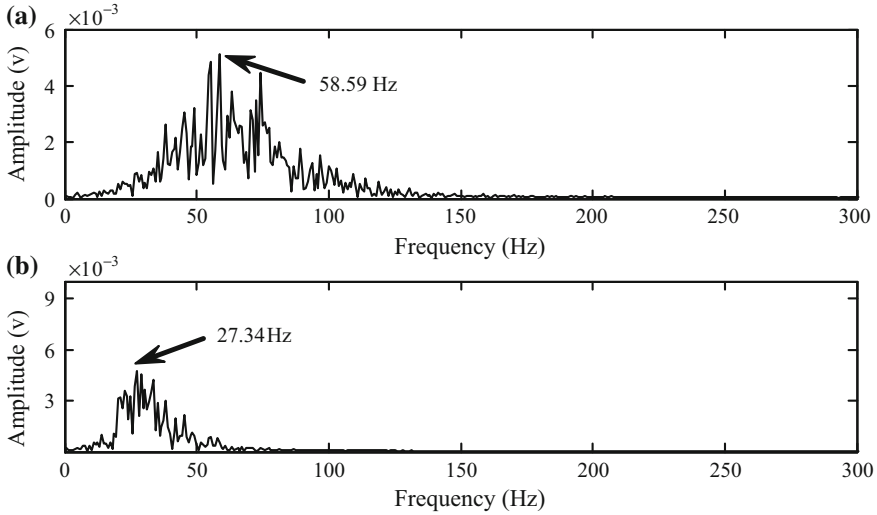


Fig. 27 **a** Fourier spectrum of the 8th IMF, and **b** Fourier spectrum of the 9th IMF using EEMD

After carefully checked the Fourier spectra of the 8th and 9th IMFs given in Fig. 27, we found two obvious peaks at the frequency of 58.59 and 27.34 Hz. These two frequencies, however, are neither the outer race fault characteristic frequency of 44.5 Hz nor the roller fault characteristic frequency of 19.94 Hz. Hence, the EEMD method fails to extract fault characteristics and diagnose the compound faults of this rolling element bearing.

For comparison, the CEEMDAN based method is applied to analyze the signal using the same noise amplitude and ensemble size in the EEMD method. There are totally 15 IMFs obtained by CEEMDAN and Fig. 28 shows the first 10 IMFs and their Fourier spectra. By examining each Fourier spectrum, we find that there are peaks at the outer race and the roller fault characteristic frequency (45.31 and 20.31 Hz) in the Fourier spectra of the 9th and 10th IMFs, shown in Fig. 29a, b, respectively. Therefore, based on the extracted fault characteristics using the CEEMDAN-based method, the compound faults of this bearing are diagnosed.

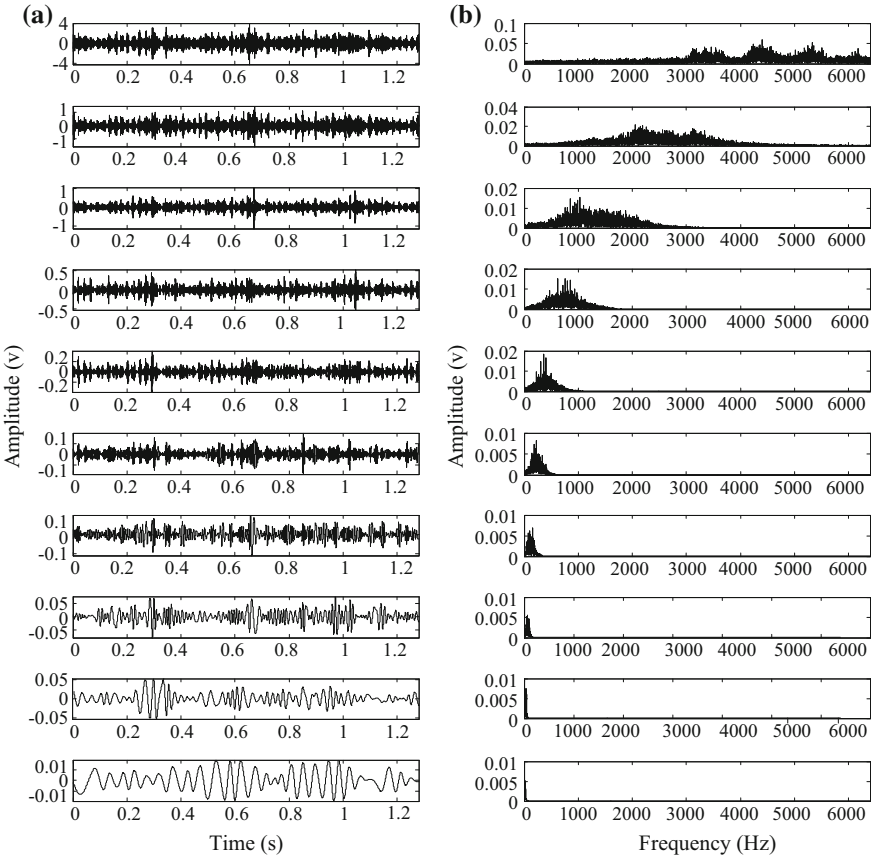


Fig. 28 Decomposition results of the vibration signal of a bearing with compound faults using the CEEMDAN method: **a** the IMFs, and **b** the Fourier spectra of IMFs

5 Conclusions

In this chapter, the basic theory of EMD and improved EMD methods, such as EEMD method, adaptive EEMD method, etc., are presented. In addition, the applications of these methods in fault diagnosis of rotating machinery, including rotors, gears and rolling element bearings, are described in details.

(a) The empirical mode decomposition (EMD) method has a good performance in the analysis of nonlinear and non-stationary signals. However, it is often subject to some problems, like end effects and mode mixing, etc. Thus the decomposed IMFs sometimes are unable to reflect the fault characteristics in fault diagnosis of rotating machinery precisely.

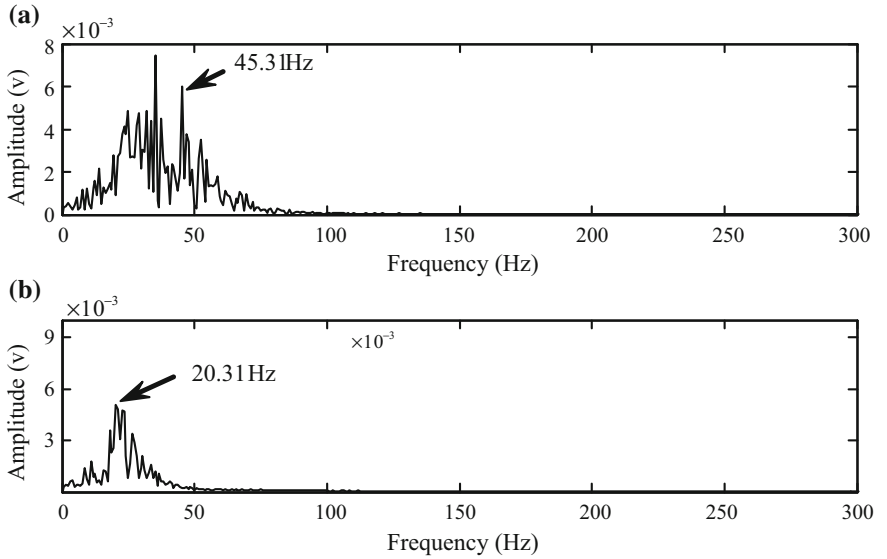


Fig. 29 **a** Fourier spectrum of the 9th IMF, and **b** Fourier spectrum of the 10th IMF using the CEEMDAN method

(b) To overcome the shortcomings of EMD, a lot of new methods based on EMD are proposed to diagnose rotating machinery improve the EMD method. These methods are developed mainly by adding different kinds of white Gaussian noise to improve the extrema distribution of the signal.
(c) The EMD and improved EMD methods have been applied in the fault diagnosis of rotating machinery. EMD sometimes cannot detect the fault in the machinery, while the improved methods promote the performance of EMD in different aspects. These methods have been proven to be effective in the diagnosis of rotors, gears and rolling element bearings.

References

1. Fan X., Zuo M.J., "Machine fault feature extraction based on intrinsic mode functions," Measurement Science and Technology, 2008, 19:045105.
2. Lei Y., He Z., Zi Y., Chen X., "New clustering algorithm-based fault diagnosis using compensation distance evaluation technique," Mechanical Systems and Signal Processing, 2008, 22:419–435.
3. Huang N.E., Shen Z., Long S.R., Wu M.C., Shih H.H., Zheng Q., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series

analysis," Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 1998, 903–995.
4. Srinivasan R., Rengaswamy R., Miller R., "A modified empirical mode decomposition (EMD) process for oscillation characterization in control loops," Control Engineering Practice, 2007, 15:1135–1148.
5. Luo L., Yan Y., Xie P., Sun J., Xu Y., Yuan J., "Hilbert–Huang transform, Hurst and chaotic analysis based flow regime identification methods for an airlift reactor," Chemical Engineering Journal, 2012, 181:570–580.
6. Xu G., Tian W., Qian L., "EMD-and SVM-based temperature drift modeling and compensation for a dynamically tuned gyroscope (DTG)," Mechanical Systems and Signal Processing, 2007, 21:3182–3188.
7. Guo Z., Zhao W., Lu H., Wang J., "Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model," Renewable Energy, 2012, 37:241–249.
8. Tang J., Zhao L., Yue H., Yu W., Chai T., "Vibration analysis based on empirical mode decomposition and partial least square," Procedia Engineering, 2011, 16: 646–652.
9. Zhang Z., Zhang Y., Zhu Y., "A new approach to analysis of surface topography," Precision Engineering, 2010, 34:807–810.
10. Charleston-Villalobos S., Gonzalez-Camarena R., Chi-Lem G., Aljama-Corrales T., "Crackle sounds analysis by empirical mode decomposition," IEEE Engineering in Medicine and Biology Magazine, 2017, 26:40–47, 2007.
11. Ambikairajah E., "Emerging features for speaker recognition," 2007 6th International Conference on Information, Communications & Signal Processing, 2007, pp. 1–7.
12. Yang Y., Chang K., "Extraction of bridge frequencies from the dynamic response of a passing vehicle enhanced by the EMD technique," Journal of Sound and Vibration, 2009, 322: 718–739.
13. Zhang H., Qi X., Sun X., Fan S., "Application of Hilbert-Huang transform to extract arrival time of ultrasonic lamb waves," International Conference on Audio, Language and Image Processing, 2008, pp. 1–4.
14. Rato R., Ortigueira M., Batista A., "On the HHT, its problems, and some solutions," Mechanical Systems and Signal Processing, 2008, 22:1374–1394.
15. Chen G., Wang Z., "A signal decomposition theorem with Hilbert transform and its application to narrowband time series with closely spaced frequency components," Mechanical Systems and Signal Processing, 2012, 28:258–279.
16. Rilling G., Flandrin P., Goncalves P., "On empirical mode decomposition and its algorithms," IEEE-EURASIP workshop on nonlinear signal and image processing, 2003, pp. 8–11.
17. Yang Z., Yang L., Qing C., Huang D., "A method to eliminate riding waves appearing in the empirical AM/FM demodulation," Digital Signal Processing, 2008, 18:488–504.
18. Tsakalozos N., Drakakis K., Rickard S., "A formal study of the nonlinearity and consistency of the empirical mode decomposition," Signal Processing, 2012, 92: 1961–1969.
19. Hawley S.D., Atlas L.E., Chizeck H.J., "Some properties of an empirical mode type signal decomposition algorithm," IEEE Signal Processing Letters, 2010, 1:24–27, 2010.
20. Wu Z., Huang N.E., "Ensemble empirical mode decomposition: a noise-assisted data analysis method," Advances in adaptive data analysis, vol. 1, pp. 1–41, 2009.
21. Lei Y., He Z., Zi Y., "Application of the EEMD method to rotor fault diagnosis of rotating machinery," Mechanical Systems and Signal Processing, 2009, 23: 1327–1338.
22. Lei Y., Zuo M., "Fault diagnosis of rotating machinery using an improved HHT based on EEMD and sensitive IMFs," Measurement Science and Technology, 2009, 20:125701.
23. Lei Y., Zuo M., Hoseini M., "The use of ensemble empirical mode decomposition to improve bispectral analysis for fault detection in rotating machinery," Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2010, 224: 1759–1769.
24. Lei Y., Li N., Lin J., Wang S., "Fault diagnosis of rotating machinery based on an adaptive ensemble empirical mode decomposition," Sensors, 2013, 13: 16950–16964.

25. Lei Y., He Z., Zi Y., "EEMD method and WNN for fault diagnosis of locomotive roller bearings," Expert Systems with Applications, 2011, 38:7334–7341.
26. Lei Y., Liu Z., Ouazri J., Lin J., "A fault diagnosis method of rolling element bearings based on CEEMDAN," Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2015, 203–210:1989–1996.
27. Flandrin P., Rilling G., Goncalves P., "Empirical mode decomposition as a filter bank," IEEE Signal Processing Letters, 2004, 11:112–114.
28. Flandrin P., Gonçalves P., Rilling G., "EMD equivalent filter banks, from interpretation to applications," The Hilbert-Huang Transform and Its Applications, edited by N. E. Huang and S. S. P. Shen, World Scientific, 2005.
29. Lei Y., Lin J., He Z., Zuo M., "A review on empirical mode decomposition in fault diagnosis of rotating machinery," Mechanical Systems and Signal Processing, 2013, 35:108–126.
30. Torres M. E., Colominas M., Schlotthauer G., Flandrin P., "A complete ensemble empirical mode decomposition with adaptive noise," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, 2011, pp. 4144–4147.
31. Colominas M.A., Schlotthauer G., Torres M. E., "Improved complete ensemble EMD: A suitable tool for biomedical signal processing," Biomedical Signal Processing and Control, 2014, 14:19–29.
32. Colominas M.A., Schlotthauer G., Torres M.E., Flandrin P., "Noise-assisted EMD methods in action," Advances in Adaptive Data Analysis, 2012, 4:1250025.
33. Lei Y., Lin J., He Z., Kong D., "A method based on multi-sensor data fusion for fault detection of planetary gearboxes," Sensors, 2012, 12:2005–2017.
34. Feng Z., Zuo M.J., "Vibration signal models for fault diagnosis of planetary gearboxes," Journal of Sound and Vibration, 2012, 331:4919–4939.
35. Samuel P.D., Pines D.J., "A review of vibration-based techniques for helicopter transmission diagnostics," Journal of Sound and Vibration, 2005, 282:475–508.

Bivariate Empirical Mode Decomposition and Its Applications in Machine Condition Monitoring

Wenxian Yang

Abstract Attributed to providing a more realistic representation of the signal without the artifacts imposed by non-adaptive limitations suffered by both Fourier- and Wavelet-transform based methods, Empirical Mode Decomposition (EMD) has been widely accepted as a favored tool for interpreting nonlinear, non-stationary signals, which are often associated with the occurrence of faults or variable operations of rotating machinery. In this chapter, the fundamental theory of the EMD will be explained. But more context will be spent on discussing its two dimensional form, namely Bivariate Empirical Mode Decomposition, and the powerful capacity of this innovative technique in the application of machine condition monitoring.

1 Introduction

Since nonlinear and/or non-stationary signal features are often associated with the occurrence of abnormalities or faults in rotating machinery, time-frequency analysis (TFA) of condition monitoring (CM) signals has become a popular approach to implement machine health assessment and thereby condition-based maintenance. Thus, an efficient and reliable TFA tool will significantly benefit the machine operation & maintenance (O&M) activities particularly in those industries highly relying on the intensive use of robotic machines.

In the past two decades, Wavelet transform (WT) has overwhelmed the traditional Fourier transform (FT) and its extension forms [e.g. Short-time Fourier Transform (STFT)] and become a favoured tool extensively applied to the TFA of machine CM signals. Thanks to the innovative multiple resolution analysis synchronously in both time and frequency domains, the invention of the WT does

W. Yang (✉)
School of Engineering, Newcastle University,
Newcastle upon Tyne NE1 7RU, UK
e-mail: wenxian.yang@ncl.ac.uk; Wenxian.Yang@newcastle.ac.uk

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors,
Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_11

significantly promote the TFA and attract lots of interests from a wide range of communities, including from machine CM researchers and engineers [1–7]. However, this does not mean the WT is non-defective. In fact, the WT has intrinsic disadvantages as well. For example, despite the unique multiple resolution capability in time and frequency domains, the WT is still not locally adaptive and unable to match the transient time-frequency features of the signals of interest, particularly those intra-wave signals collected from variable speed machines. For this reason, the WT sometimes cannot provide an accurate representation of the CM signals, although it does show powerful ability in feature extraction in most cases [8]. Moreover, the DC component contained in the CM signal still preserves in the WT results, which will affect the identification of the fault-related features in grayscale time-frequency map of the CM signal where the energies of frequency components are indicated using various colours [9]. As a consequence, the WT meets difficulty sometimes especially when the inspected fault is at its early developing stage due to the resultant 'overlaps' in frequency vicinity and the weak fault-related features. The worse thing is that it has no way to remove these issues because they are inborn with the WT. This motivates the research on an alternative TFA method, namely Empirical Mode Decomposition (EMD) [10].

2 Empirical Mode Decomposition

Different from the conventional TFA methods, the EMD is completely a data driven techniques. Through automatically performing a series of recursive calculations, it decompose the signal of interest automatically into a finite number of basic oscillation modes, namely Intrinsic Mode Functions (IMFs). Then, the transient time-frequency features of the inspected signal can be extracted out by interpreting these IMFs using Hilbert transform. This is the so called Hilbert-Huang Transform (HHT) [11]. Assume a signal $x(t)$, its EMD can be implemented by following the steps depicted below [8, 10].

Step 1. Initialise $r_0 = x(t)$ and $i = 1$;
Step 2. Extract the i-th IMF by conducting the following recursive calculations.

(1) Initialise $h_{i(k-1)} = r_i$, $k = 1$;
(2) Extract the local extrema and minima of $h_{i(k-1)}$;
(3) Interpolate the local extrema and the minima by cubic spline lines to form the upper and lower envelops of $h_{i(k-1)}$;
(4) Calculate the mean of the obtained upper and lower envelops and denote it by $m_{i(k-1)}$;
(5) Let $h_{ik} = h_{i(k-1)} - m_{i(k-1)}$;
(6) Check whether the resultant h_{ik} satisfies the following two conditions.

(a) In the whole data set, the number of extrema and the number of zero crossings must either equal or differ at most by one; and
(b) At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

If h_{ik} satisfies the above two conditions, it is an IMF. Otherwise, it is not an IMF;

(7) If h_{ik} is an IMF, let $IMF_i = h_{ik}$, else go to (2) to repeat the calculations with $k = k + 1$ until an IMF_i is achieved successfully;

Step 3. Calculate $r_{i+1} = r_i - IMF_i$;

Step 4. If the obtained r_{i+1} still has over 2 extrema, go to Step 2 to iterate the decomposition to get more IMFs, else finish the EMD calculation and define r_{i+1} as the residue of the signal.

In the end, the original $x(t)$ can be represented by the sum of a collection of IMFs and the residue, i.e.

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

where r_n is the final residue, n indicates the number of IMFs obtained.

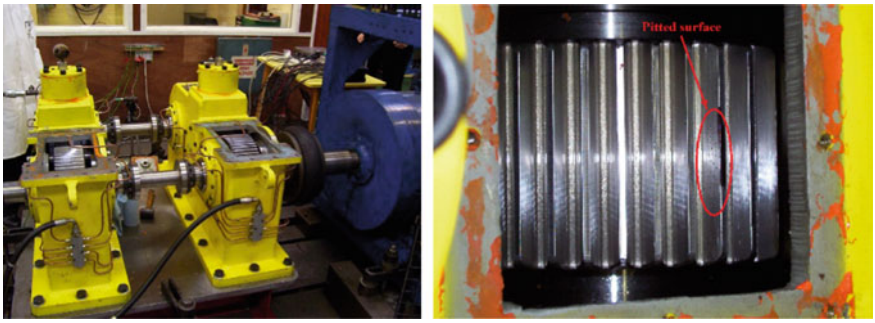
From the above descriptions, it is found that the EMD works directly on the time waveform of the signal and moreover without doing any assumption of the basic oscillation mode of the signal. So, it is locally adaptive and thus able to provide a 'real-life' representation to the signal of interest. This fully overcomes the disadvantages induced by the artefact imposed by the artificial assumptions and therefore non-adaptive limitations of both the FT and WT.

Attributed to the locally adaptive feature, the EMD is potentially a powerful tool for interpreting nonlinear and non-stationary signals. Nowadays, the EMD has been extensively used for detecting the faults occurring in machines. The following is an example of its application in the field of machine CM.

A defective gearbox is considered in this example. The surface of one gear of this gearbox is pitted due to overloading. The defective gear is shown in Fig. 1.

In the experiment, the vibration acceleration signals were collected by using a sampling frequency of 20 kHz. The accelerometer for data acquisition was mounted on the case of the gearbox. To facilitate understanding, the vibration signals collected before and after the gear surface was pitted are shown in Fig. 2. Where, the first 0.4 s data is the signal for healthy gearbox, while the second 0.4 s is the signal obtained after the gear became defective.

Then, apply the EMD to analysing the signal. The corresponding signal decomposition results are shown in Fig. 3.



(a) Gearbox (b) Defective gear surface

Fig. 1 Gearbox being inspected

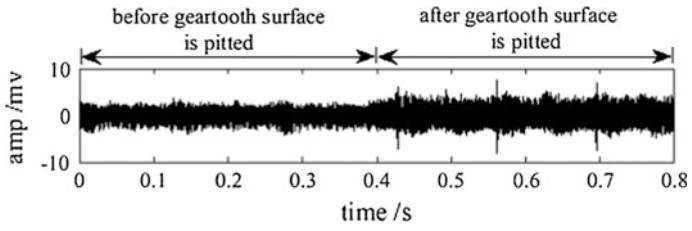


Fig. 2 Gearbox vibration signal

From Fig. 3, it can be found that the raw signal is decomposed into 13 IMFs. The characteristic impact features induced by the pitted gear surface can be clearly observed from the first IMF, while they are absent from the time waveform of 'imf1' when the gearbox is healthy. Moreover, Fig. 3 shows that all obtained IMFs are zero-mean oscillations. Therefore, the negative influence of variable operational conditions of the machine on CM is not an issue for the EMD-based CM technique. Thus, through observing the time waveforms of the EMD resultant IMFs, the health condition of the machine may be correctly assessed. Moreover, as the EMD does not involve any complex transform calculations like the WT does, the EMD-based CM method is more efficient in computation.

However, the EMD was initially designed to processing one-dimensional signals. It is unable to execute information fusion, which considers multiple CM signals collected simultaneously from the inspected machine and has potential to lead to more reliable CM conclusion. Accordingly, to enhance the capability of the EMD, Complex Empirical Mode Decomposition [12], also namely Bivariate Empirical Mode Decomposition (BEMD), was proposed [13].

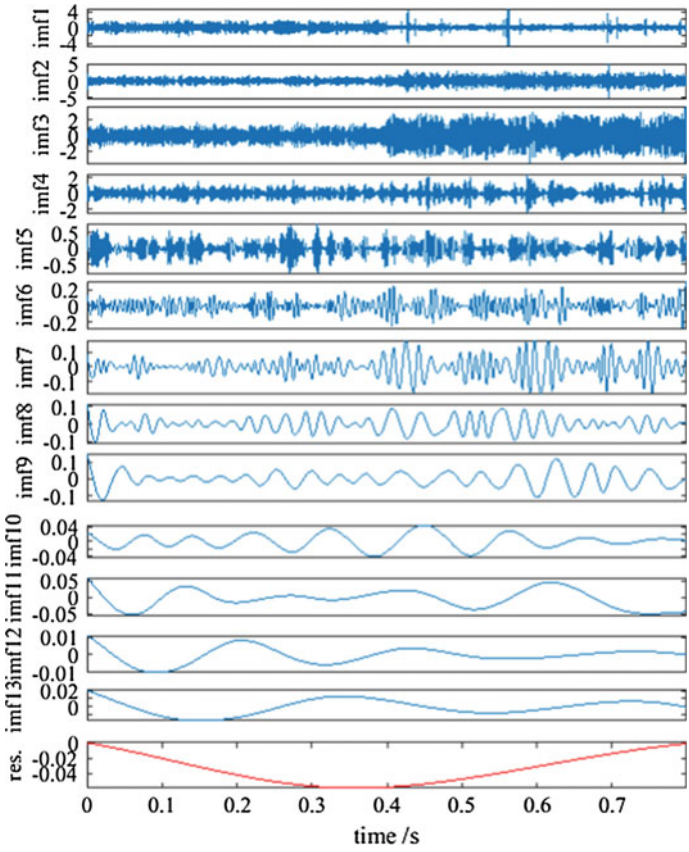


Fig. 3 The EMD results of the gearbox vibration signal

3 Bivariate Empirical Mode Decomposition

The BEMD was designed specifically to deal with complex-valued signals. In the application of rotating machine CM, the complex-valued signal can be constructed by using the vibration signals collected in two mutually perpendicular directions. But in the CM of other machinery, specific conversions are often required when constructing the complex-valued signal. An example of this conversion will be given in next Section to ease understanding.

The calculation method of the BEMD is similar as that of the EMD except some necessary modifications in extrema detection and envelope definition. Assume a complex valued signal $x(t)$, the computing algorithm of its BEMD is described as follows [13].

Step 1. Determine the number of projection directions N and calculate the projection directions φ_n, i.e.

$$\varphi_{\mathrm{n}} = \frac{2\pi}{\mathrm{N}} \times \mathrm{n} \quad \mathrm{n} \in [1, \mathrm{N}] \tag{2}$$

Step 2. Project the complex valued signal $\mathrm{x(t)}$ on the direction φ_{n} by

$$P_{\varphi_n}(t) = \mathrm{Re}\left(\mathrm{e}^{-j\varphi_n} \times x(t)\right) \quad j = \sqrt{-1} \tag{3}$$

Step 3. Find all local maxima of $P_{\varphi_n}(t)$ and record their locations and values $\left\{\left(t_i^n, P_{\varphi_n}\left(t_i^n\right)\right)\right\}$. Herein, i indicates the No. of individual local maxima;

Step 4. Interpolate the set $\left\{\left(t_i^n, P_{\varphi_n}\left(t_i^n\right)\right)\right\}$ by using cubic spline line to obtain the envelop curve $e_{\varphi_n}(t)$ in direction φ_n;

Step 5. Repeat Steps 2–4 until the envelop curves in all N projection directions are obtained;

Step 6. Compute the mean of all envelop curves by

$$m(t) = \frac{1}{N}\sum_{n=1}^{N} e_{\varphi_n}(t) \tag{4}$$

Step 7. Subtract $m(t)$ from $x(t)$ and obtain $h(t)$, i.e.

$$h(t) = x(t) - m(t) \tag{5}$$

Step 8. Examine whether the obtained $h(t)$ satisfies the conditions of an IMF. If not, let $x(t) = h(t)$ and repeat the calculations depicted from Steps 2–7 until the obtained $h(t)$ is an IMF. If yes, go to next step;

Step 9. Record the obtained $h(t)$ as an IMF and remove it from original signal $x(t)$, i.e.

$$\mathrm{IMF}_1(t) = h(t) \tag{6}$$

$$r_1(t) = x(t) - \mathrm{IMF}_1(t) \tag{7}$$

Step 10. Treat $r_1(t)$ as a new original signal and repeat above calculations until achieve a new IMF. Then, calculate the new residual component by

$$r_2(t) = r_1(t) - \mathrm{IMF}_2(t) \tag{8}$$

Step 11. Iterate the above calculations until all IMFs embedded in the original signal $x(t)$ are obtained. Finally, the signal $x(t)$ can be expressed as

$$x(t) = \sum_{k=1}^{K} \mathrm{IMF}_k(t) + r_k(t) \tag{9}$$

where K indicates the total number of the IMFs extracted from the original signal.

From the computing algorithm described above, it is found that the BEMD performs the decomposition of the real and imaginary parts of the complex-valued signal simultaneously. Such a calculation method not only significantly improves the computing efficiency, but also perfectly preserves the phase information of the signal in the decomposition results. All these advantages will benefit machine CM. Some examples are given below for demonstration.

4 Applications of the BEMD in Machine Condition Monitoring

In this section, the superiorities of the BEMD over the traditional EMD in the field of machine CM will be demonstrated via a few examples.

4.1 The CM of Bearing-Shaft Systems

The bearing-shaft systems have been extensively used in rotating machinery. They are critical subassemblies to assure the machines are able to successfully deliver assigned tasks. So, it is necessary to understand their instant operational and health conditions via a reliable CM system. Since allowing the fusion of three dimensional information, the BEMD is regarded as one of the most promising techniques for machine CM. Such point of view will be demonstrated by the CM practice depicted below.

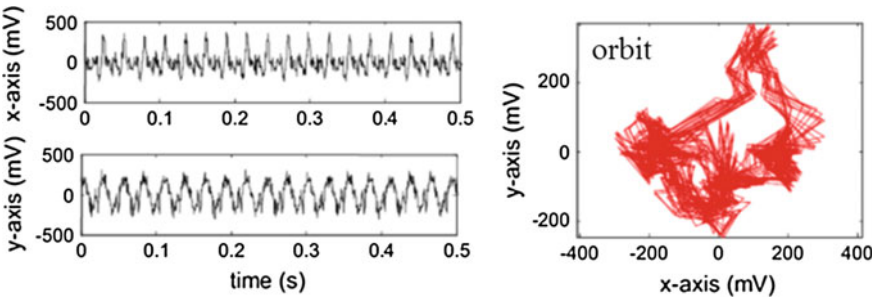


Fig. 4 The shaft vibration signals from a compressor with rotor-stator rubbing fault [15]

(a) Rotor-stator rubbing

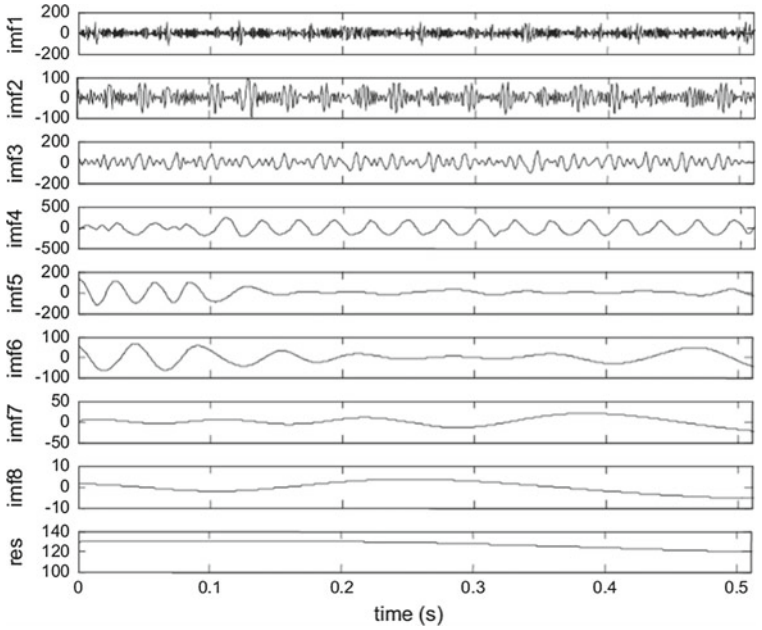
Firstly, the shaft vibration signals collected from a centrifugal compressor with rotor-stator rubbing fault are investigated by using the BEMD. The signals were collected by using a sampling frequency of 2 kHz from two mutually perpendicular proximeters installed on the bearing case. The obtained vibration signals and the corresponding shaft vibration orbit are shown in Fig. 4.

In [14], both the vibration signals shown in Fig. 4 were processed by using the EMD in order to extract the purified vibration orbit of the shaft, so that the rotor-stator rubbing phenomenon can be observed clearly from the purified pattern. This is indeed an innovative idea to reconstruct the vibration orbit of a shaft-bearing system by taking advantage of the EMD in signal representation. However, the separate decomposition of the signals collected in two directions by using the EMD cannot preserve the relative phase information between two signals. In addition, owing to the EMD is very sensitive to the noise contained in the signals, different numbers of the IMFs might be obtained from the two signals if they are respectively processed by the EMD, as shown in Fig. 5. Then, how to select appropriate IMFs to reconstruct the purified shaft vibration orbit will become a difficult issue. Due to these limitations, the purified shaft orbit obtained by the traditional EMD approach might be unable to correctly reflect the actual machine vibration and probably lead to wrong machine CM conclusion. In view of this, the BEMD is applied to interpret the signals in [15].

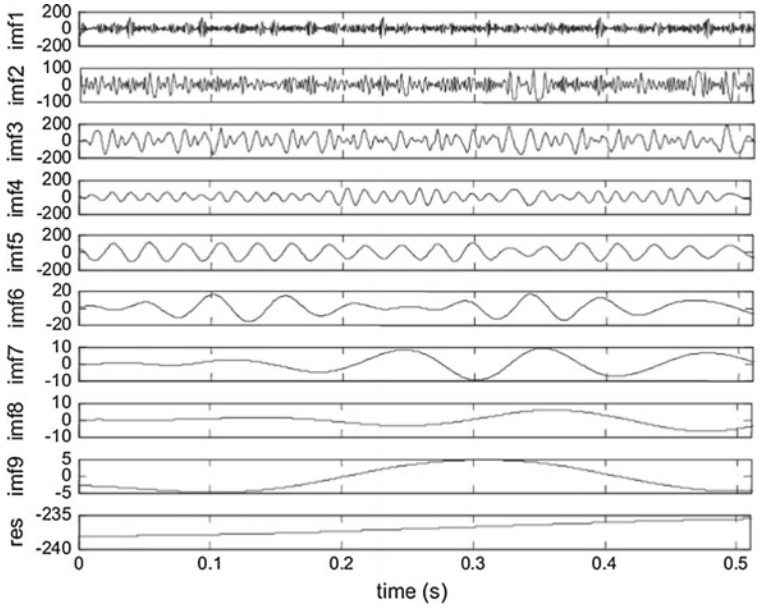
As mentioned in Sect. 3, use the vibration signals collected in two mutually perpendicular directions $x(t)$ and $y(t)$ to construct a new complex-valued signal $z(t)$, i.e.

$$z(t) = x(t) + jy(t), \quad j = \sqrt{-1} \tag{10}$$

Apply the BEMD computing algorithm depicted in Sect. 3 to $z(t)$, the corresponding signal decomposition results are shown in Fig. 6. Where, the black lines represent the real parts while the red lines are the imaginary parts of the complex-values IMFs.



(a) x-direction



(b) y-direction

Fig. 5 The EMD results for the shaft vibration signals in Fig. 4

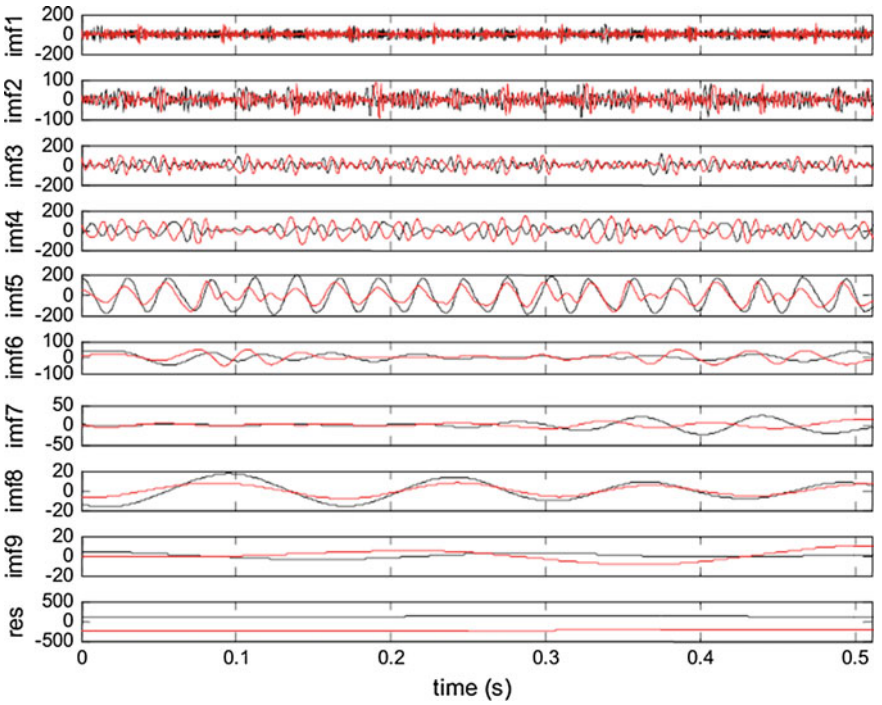


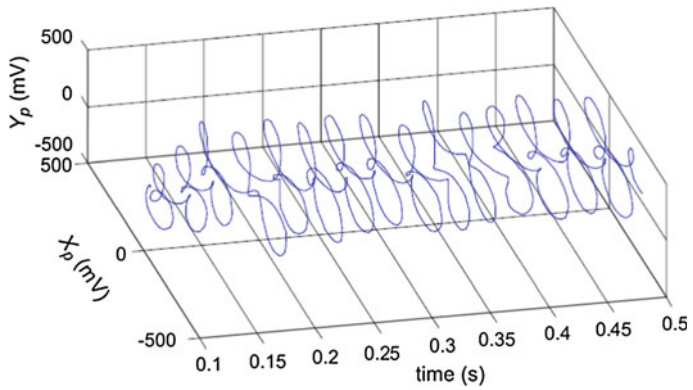
Fig. 6 The corresponding BEMD results for the shaft vibration signals in Fig. 4 [15]

Since in the BEMD calculations, the signals $x(t)$ and $y(t)$ are treated as one signal and decomposed simultaneously, not only the computational efficiency is improved, but also the phase information between signals $x(t)$ and $y(t)$ is perfectly preserved in the BEMD results. Moreover, they are decomposed into the same number of IMFs, avoiding the orbit construction difficulties met in the EMD scenario. All these merits of the BEMD are very helpful to improve the reliability of machine CM results. For comparison, the three dimensional purified shaft vibration orbits obtained respectively by the EMD and BEMD approaches are shown in Fig. 7. Both are reconstructed by using the first six IMFs.

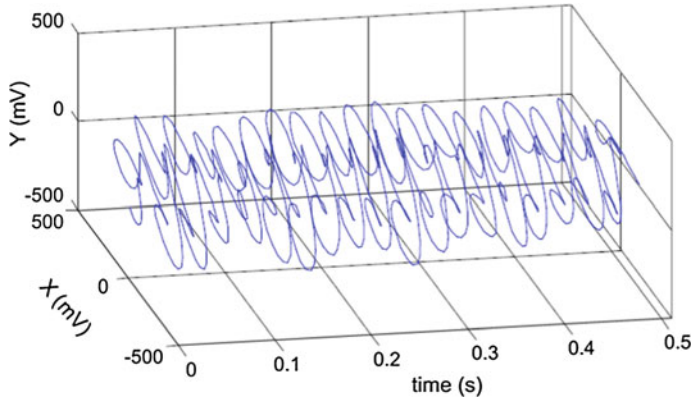
From Fig. 7, it is clearly seen that, in comparison of the result by the EMD, the purified shaft vibration orbit obtained by the approach of the BEMD better reveals the sharp turning behavior of the shaft due to rotor-stator rubbing.

(b) Fluid excitation

Fluid excitation is a phenomenon often occurs in large steam turbines or compressors. Once happened, they will cause significant vibration to machine structures thus result in additional fatigue issues, reducing the reliability of the machine and even causing immediate damage to the machine structure in worse case. Figure 8



(a) The result by the EMD



(b) The result by the BEMD

Fig. 7 The purified shaft vibration orbit

shows the vibration displacement signals collected from a compressor when it was experiencing fluid excitation. The signals are from two transducers that are mounted in two mutually perpendicular directions.

From Fig. 8, it is seen that both signals are nonlinear and non-stationary over time. So, strictly speaking, the traditional FT-based techniques are not suitable to be applied to process this kind of signals, although the FT and its extension forms have been extensively used in the past decades. In order to diagnose this fault, both two dimensional and the corresponding three dimensional shaft vibration orbits are drawn in Fig. 9, so that more shaft vibration information can be extracted from them.

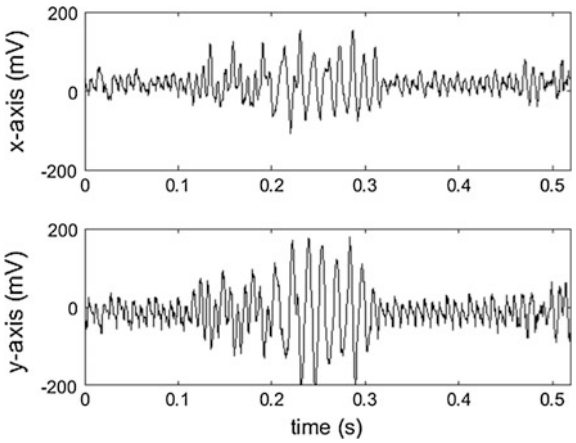
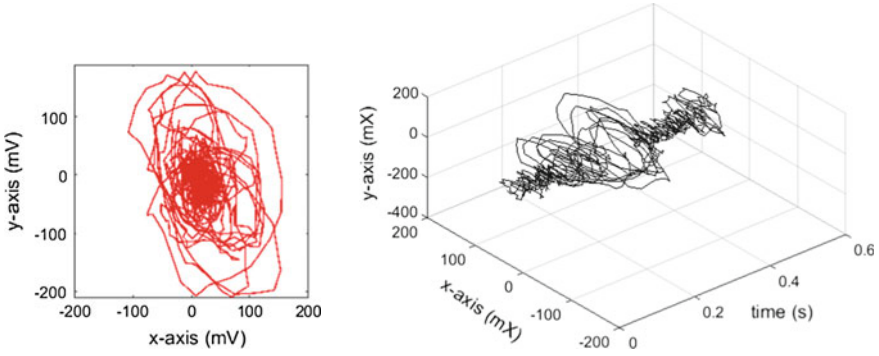


Fig. 8 Vibration signals when the machine experience fluid excitation



(a) Two dimensional orbit (b) Three dimensional orbit

Fig. 9 Shaft vibration orbit when fluid excitation happens

From Fig. 9, it is found that in contrast to the two dimensional orbit shown in Fig. 9a, the three dimensional orbit in Fig. 9b obviously discloses more details about the shaft vibration. The fluid excitation induced nonlinear and non-stationary vibration of the shaft can be observed more clearly. However, owing to the influence of background 'noise' contained in the signals, the nonlinear and non-stationary vibration behavior of the shaft is still not vividly displayed in Fig. 9b. Therefore, the BEMD is adopted to purify the signals. The BEMD results and the corresponding purified three dimensional shaft vibration orbit are shown in Fig. 10. Where, the purified shaft vibration signals are reconstructed by using the third and forth IMFs (i.e. imf3 and imf4) as they are large in amplitude thus dominate the vibration of the shaft.

From Fig. 10, it is obviously found that the fluid excitation induced nonlinear and non-stationary shaft vibration features have been vividly illustrated from both purified signals and the resultant three dimensional shaft vibration orbit. This will be very helpful to the operator to obtain right assessment of the machine operational and health condition.

(c) Looseness

Owing to the poor installation and constant vibration, component looseness could happen in all kind of machines, including rotating machinery. If it cannot be detected, the operational performance of the machine will be affected. In the worst case, it could lead to catastrophic damage to the machine. Therefore, the instant detection of component looseness fault is helpful not only to ensure the operating performance of the machine, but also to assure its safety.

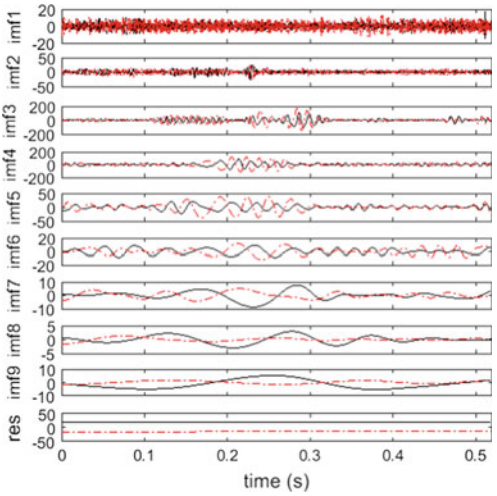
In order to demonstrate the contribution of the BEMD to the detection of this kind of faults, Fig. 11 shows two shaft vibration signals collected from a rotating machine when the machine suffers component looseness issue. The signals were collected also from two proximeters installed in mutually perpendicular directions.

To facilitate analysis, the two dimensional and three dimensional shaft vibration orbits are also considered. They are shown in Fig. 12.

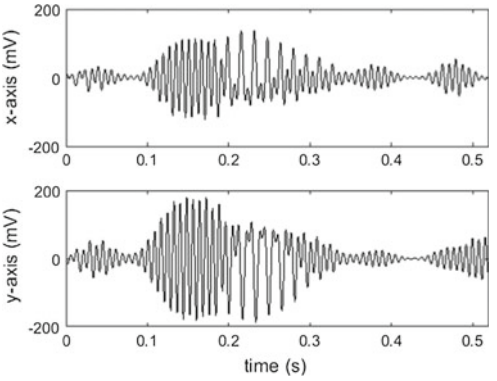
From the vibration signals and shaft vibration orbits shown in Figs. 11 and 12, it can be sure that the signals are nonlinear and non-stationary over time. However, it is difficult to relate them to the shaft vibration of a rotating machine. Despite of the types of the faults, the shaft vibration signals are always associated with the harmonic oscillations and their compositions by various means. But such an empirical knowledge can hardly be demonstrated via the signals shown in Figs. 11 and 12. This could confuse the operator and even mislead them to reach a wrong CM conclusion. Therefore, the BEMD is tried in the following to see whether it can help to clarify the misunderstanding. Accordingly, the BEMD is applied to analyze the complex-valued signal that is constructed by using the two raw vibration signals in Fig. 11. The BEMD results, the resultant purified shaft vibration signals, and the corresponding purified shaft vibration orbit are shown in Fig. 13. Where, the purified shaft vibration signals are reconstructed by using the first four IMFs (i.e. imf1, imf2, imf3 and imf4) as they are large in amplitude and dominate the vibration of the shaft.

From Fig. 13, it has no doubt that the looseness induced nonlinear and non-stationary shaft vibration has been vividly represented by the purified signals and the corresponding shaft vibration orbit.

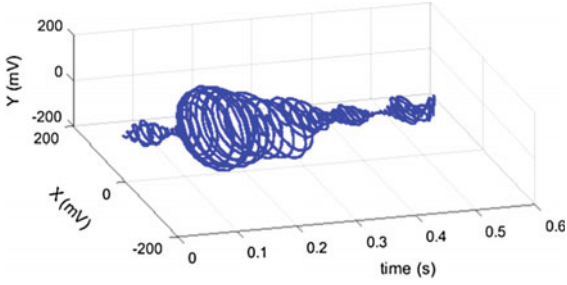
From the three examples depicted above, it can be concluded that the BEMD does show great advantages in both signal decomposition and signal presentation. In other words, the BEMD perfectly preserves the phase information of vibration signals, which guarantee the correctness of the reconstructed signals. Attributed to this merit, the BEMD allows a vivid description of the shaft vibration behavior, which is of great importance to realizing the reliable CM of rotating machinery.



(a) The BEMD results



(b) Purified shaft vibration signals



(c) Purified shaft vibration orbit

Fig. 10 The detection of fluid excitation fault by the BEMD

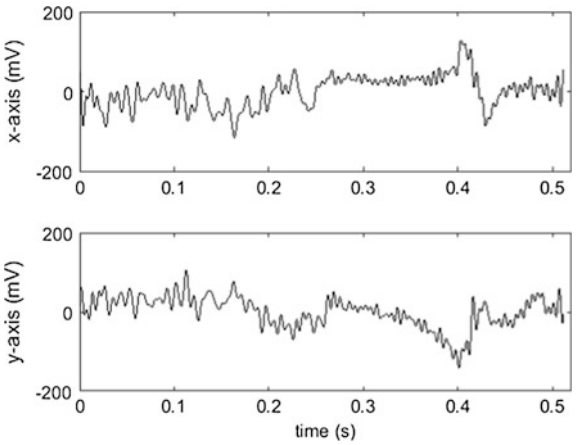
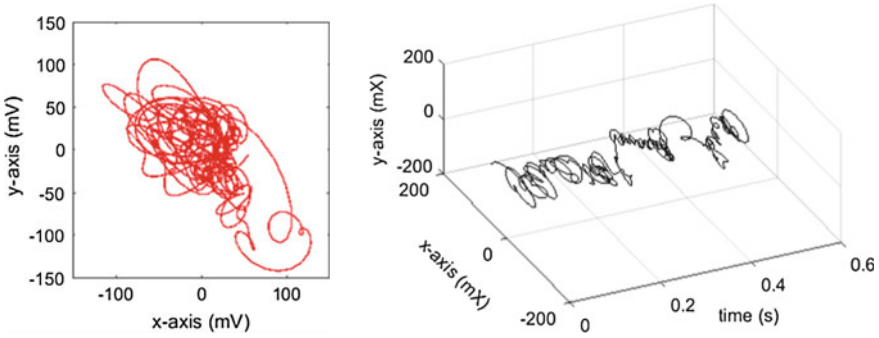


Fig. 11 Vibration signals when looseness happens

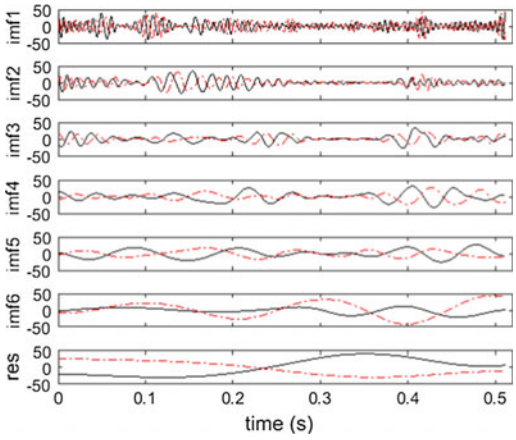


(a) Two dimensional orbit (b) Three dimensional orbit

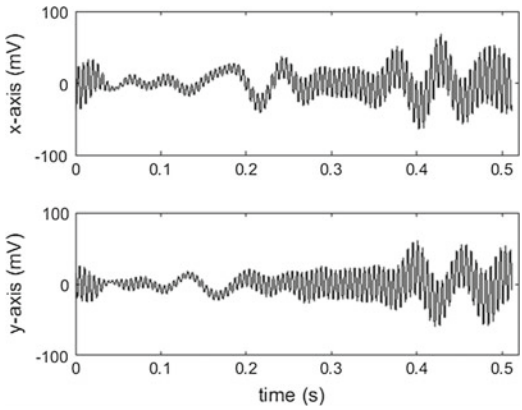
Fig. 12 Shaft vibration orbit in the presence of looseness

4.2 The CM of Wind Turbines

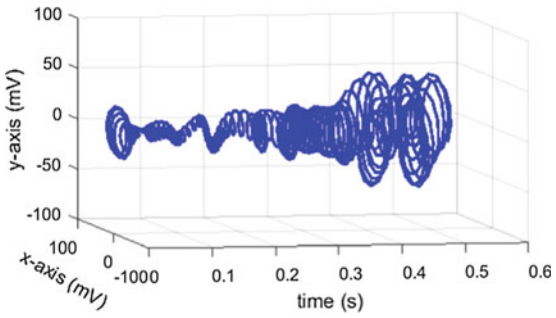
The application of the BEMD was also extended to the CM of variable speed machinery, for example wind turbines. Since wind turbines operate directly in harsh environment and are subjected to constantly varying loads, they suffer a number of reliability issues. Thus, it is very necessary to monitor them appropriately. However, as wind turbines always operate at variable speeds in order to capture more energy from wind, the varying operational and loading conditions make the wind turbine CM signals more complex in both time and frequency domains. So, the CM of wind turbine is more difficult and challenging [16].



(a) The BEMD results



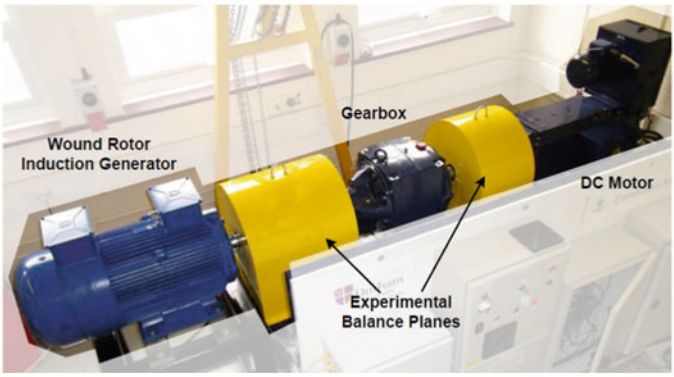
(b) The purified shaft vibration signals



(c) The purified shaft vibration orbit

Fig. 13 The detection of component looseness fault

(a) Side view



(b) Plan view

Fig. 14 Wind turbine CM test rig

To develop reliable wind turbine CM techniques, a wind turbine CM test rig was specifically developed at the University of Durham, as shown in Fig. 14.

The test rig comprises a 54 kW DC variable-speed motor, a two-stage gearbox and a 30 kW three-phase four-pole wound-rotor induction generator, instrumented and controlled using LabVIEW. In the experiments, a variety of wind speed inputs could be applied to the test rig via the DC motor, the speed of which is controlled by an external model incorporating the properties of natural wind at a variety of speeds and turbulences and the mechanical behavior of a 2 MW wind turbine operating under closed-loop conditions. Relevant CM signals were collected from the terminals of the generator and the drive train when subjected to this driving speed.

Herein, it is worth to note that the stator of the induction generator fed the three-phase mains, while its rotor circuit is coupled via slip rings to an external three-phase resistive load bank, so that the rotor electrical imbalance fault can be applied to the test rig. In addition, a circular plate was mounted on the input shaft of

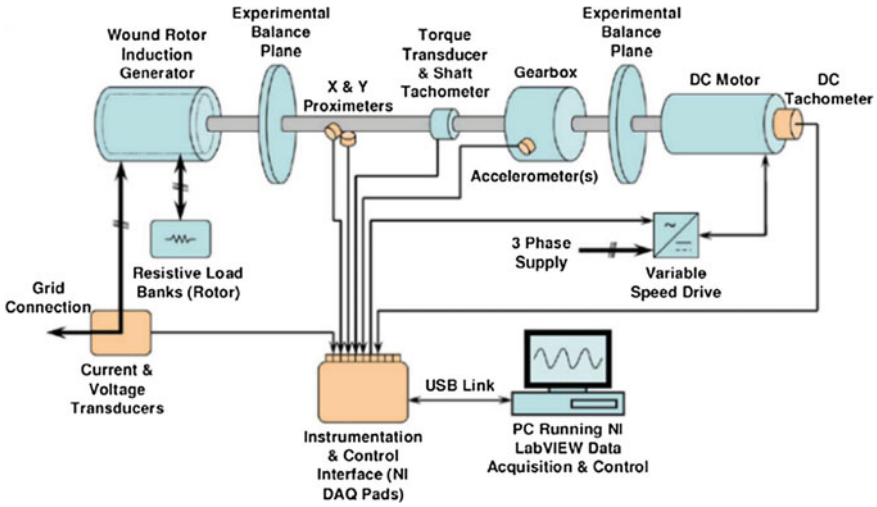


Fig. 15 The schematic diagram of the test rig [18]

the generator, on which various sizes of unbalance masses are allowed to install to emulate the mechanical unbalance fault. To facilitate understanding, the schematic diagram of the test rig is illustrated in Fig. 15.

(a) Electrical fault detection

Firstly, different severity levels of 'rotor winding faults' were emulated on the test rig by repeatedly adjusting the phase resistances of the generator rotor with the aid of the external resistive load bank. The obtained electrical current and voltage signals are shown in Fig. 16.

As shown in Fig. 16, in total 3 line electrical (I_R, I_Y, I_B) and 3 phase voltage signals (V_R, V_Y, V_B) were collected in the experiment. Considering the EMD cannot process multiple signals in parallel, the electric power based on these current and voltage signals is calculated. The obtained power signals and its EMD results are shown in Fig. 17.

From the EMD results shown in Fig. 17, it is difficult to find any convincing evidence that can indicate the rotor winding faults and their severity levels. From the time waveform of the forth IMF (i.e. imf4), the serious winding fault can be perceived, but it is still not convincing enough as a proof of CM. Therefore, the BEMD is applied to analyze the signals. But as mentioned above, a complex-valued signal should be constructed before the application of the BEMD.

In order to convert these electric current and three voltage signals to be a complex-values signal $z(t)$ that can be processed by the BEMD, the following conversion is performed based on the idea of Parker's vector [17].

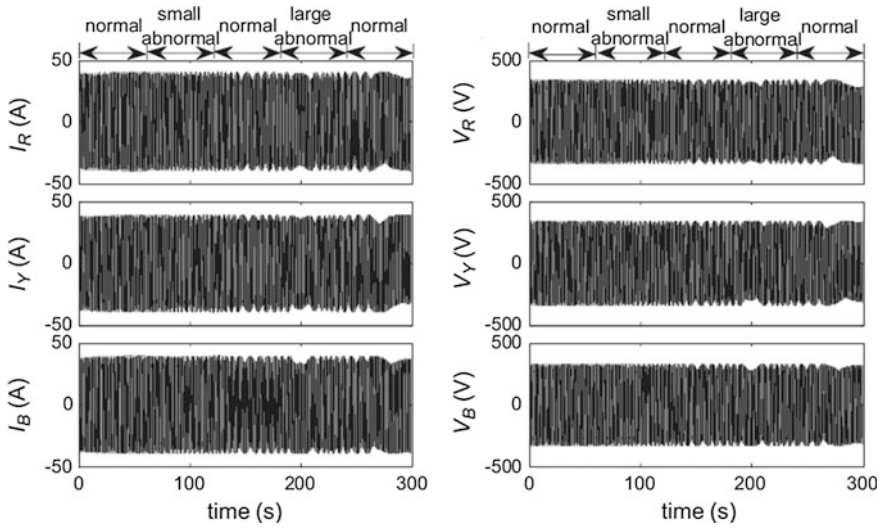


Fig. 16 Electrical signals collected from the wind turbine CM test rig [15]

$$\begin{bmatrix} I_\alpha \\ I_\beta \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} I_R \\ I_Y \\ I_B \end{bmatrix} \tag{11}$$

$$\begin{bmatrix} V_\alpha \\ V_\beta \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} V_R \\ V_Y \\ V_B \end{bmatrix} \tag{12}$$

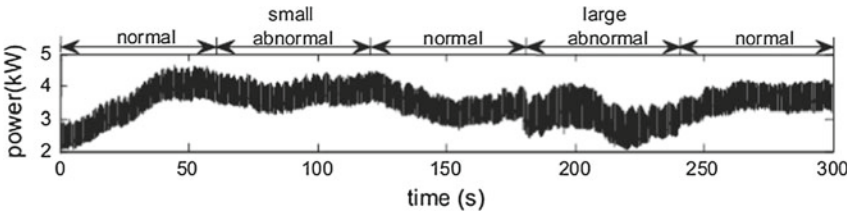
$$\begin{bmatrix} P \\ Q \end{bmatrix} = \begin{bmatrix} V_\alpha & V_\beta \\ -V_\beta & V_\alpha \end{bmatrix} \begin{bmatrix} I_\alpha \\ I_\beta \end{bmatrix} \tag{13}$$

Then, a complex-valued signal can be constructed from the resultant $P(t)$ and $Q(t)$, i.e.

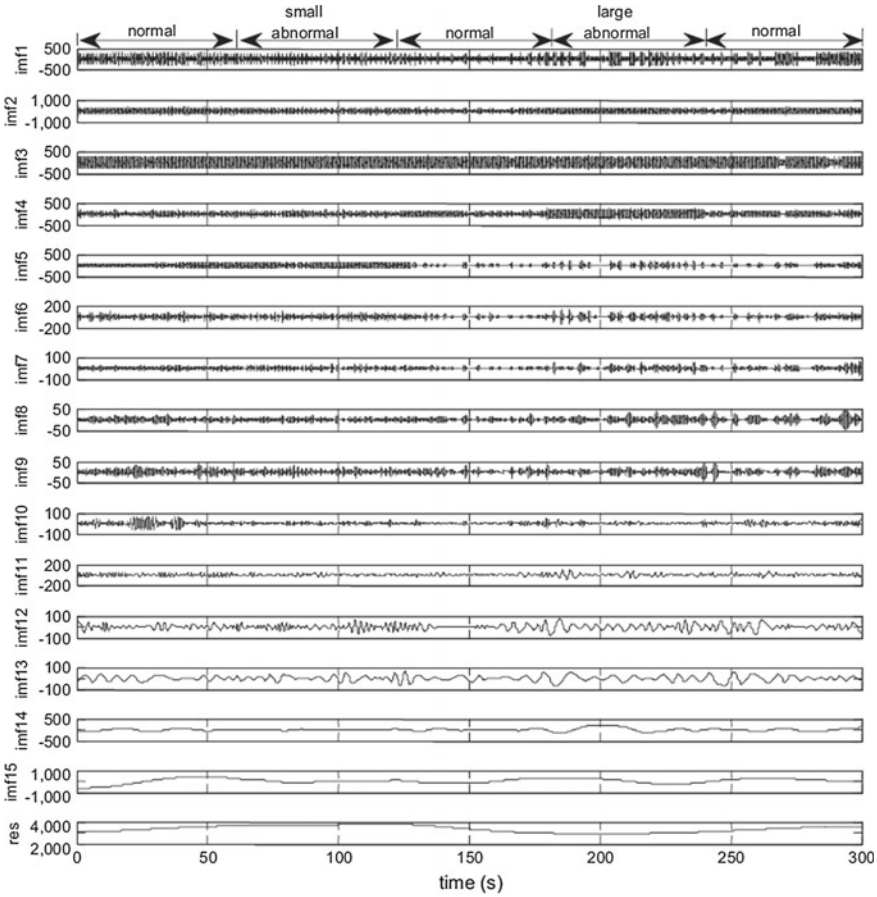
$$z(t) = P(t) + jQ(t), \quad j = \sqrt{-1}, \tag{14}$$

Apply the BEMD to the obtained $z(t)$ and the corresponding results are shown in Fig. 18.

From Fig. 18, it is found that both the real and imaginary parts of the third IMF (i.e. imf3) show changes when the electric asymmetry was applied to the rotor of the generator. Thus, imf3 can be regarded as the purified signal to conduct further CM. The amplified time waveforms of the real and imaginary parts of imf3 are illustrated in Fig. 19 to ease understanding.



(a) Electric power signal



(b) The EMD result

Fig. 17 The electric power signals and its EMD results [15]. **a** Electric power signal [15]. **b** The EMD result

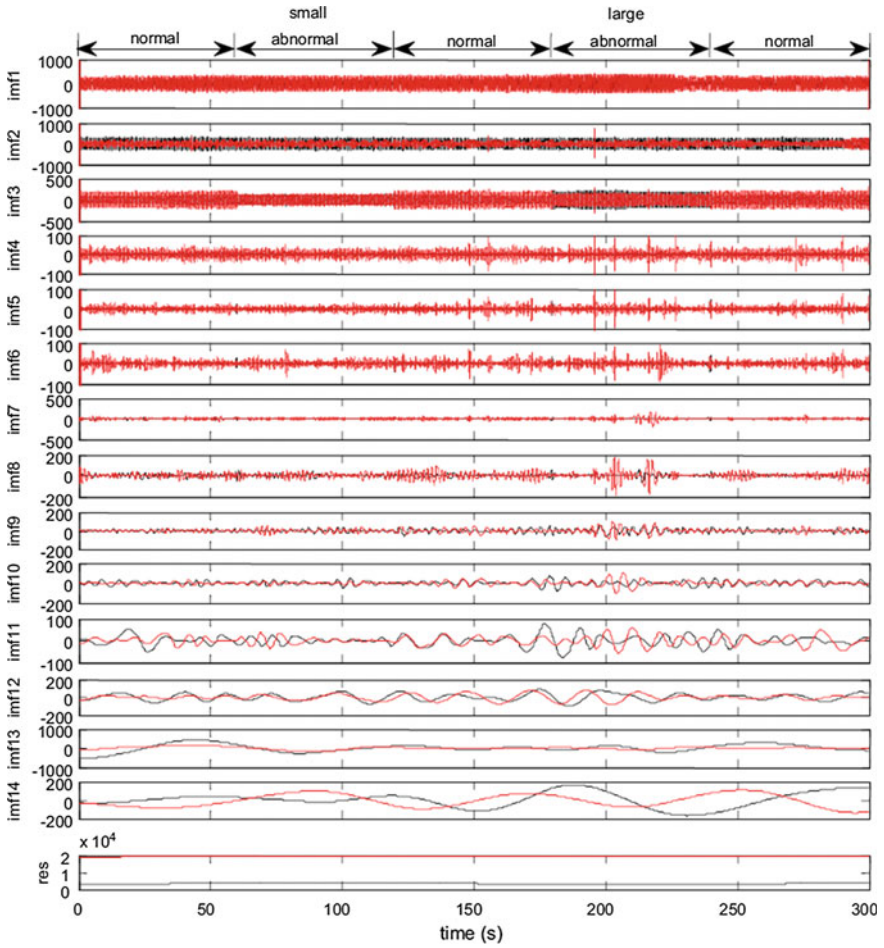


Fig. 18 The BEMD results of the complex-valued electrical signal [15]

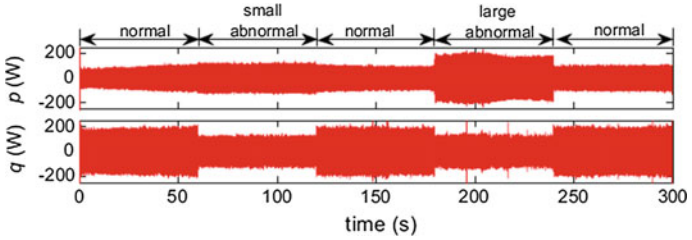


Fig. 19 Purified electrical signals by the BEMD [15]

However, the fault severity levels still cannot be identified from the purified signals shown in Fig. 19. Thus, the following CM criterion is employed. The corresponding calculation result is shown in Fig. 20.

$$\lambda(t) = \sqrt{\frac{\sum_{t=0}^{T}[p(t) - \bar{p}]^2}{\sum_{t=0}^{T}[q(t) - \bar{q}]^2}} \tag{15}$$

where $\bar{p}$ and $\bar{q}$ represent the average values of the two purified signals $p(t)$ and $q(t)$.

From Fig. 20, it is interestingly found that the severity levels of the faults have been correctly identified, i.e. the slight fault is indicated by a smaller value of λ, while the serious fault is indicated by a larger value of λ.

(b) Mechanical fault detection

Then, a rotor mechanical unbalance fault was emulated on the wind turbine CM test rig by applying an unbalance mass to the circular plate. It has been shown that the generator electrical signals are also sensitive to the mechanical faults occurring in wind turbine drive train as these faults will either disturb the torque transmission by the drive train or modify the electromagnetic field within the rotor-stator gap of the generator. However, the mechanical vibration signals are insensitive to the electrical faults occurring in wind turbine generator. For this reason, the electrical current and voltage signals were collected from the terminals of the generator before and after the mechanical unbalance fault was applied. The corresponding signal waveforms are shown in Fig. 21.

Likewise, to highlight the outstanding capability of the BEMD in both signal decomposition and signal presentation and therefore its unique contribution to machine CM, the EMD is employed first for comparison. Considering the EMD is only able to process one dimensional signals, the electric power signal was calculated from the signals shown in Fig. 21. The time waveform of the resultant electrical power signal and its EMD results are shown in Fig. 22.

From Fig. 22, it is seen that only the first IMF (i.e. imf1) derived by the EMD seems indicate the presence of the 'mechanical unbalance fault'. But the

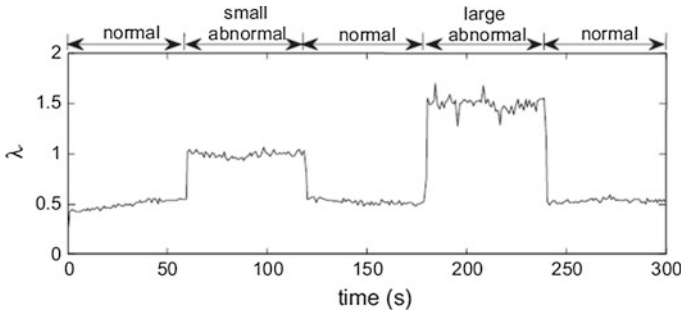


Fig. 20 CM result obtained by the approach of the BEMD [15]

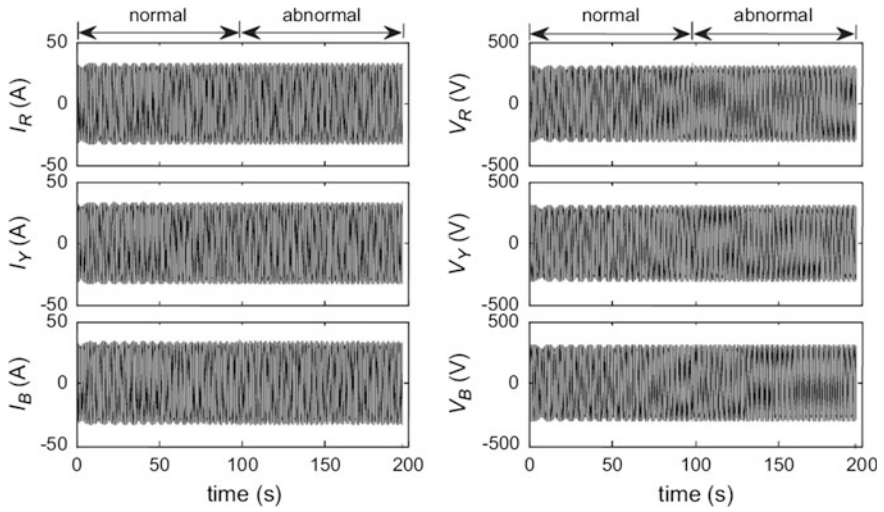


Fig. 21 Electrical signals obtained when emulating mechanical unbalance fault [15]

fault-induced change in signal amplitude is too small to be accepted as a convincing CM proof. Thus, the BEMD is adopted for detecting the fault.

As done above, to enable the BEMD analysis the electric current and voltage signals shown in Fig. 21 are converted to be a complex-valued signal first by using the Eqs. (11)–(14), and then the BEMD is applied to the resultant complex-valued signal. The corresponding signal decomposition results are shown in Fig. 23.

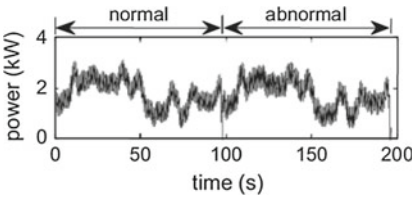
From Fig. 23, it is seen that in contrast to the EMD results, both the real and imaginary parts of the first and second IMFs (i.e. imf1 and imf2) obtained by the BEMD have presented much clearer indication to the presence of the 'mechanical unbalance fault'. Hence, the purified real and imaginary parts, i.e. the purified $p(t)$ and $q(t)$, of the signal are reconstructed by using imf1 and imf2. The resultant purified signals are shown in Fig. 24.

From Fig. 24, it is found that both the purified $p(t)$ and $q(t)$ do provide clear indication to the 'mechanical unbalance fault'. Furthermore, the CM criterion depicted by Eq. (15) is calculated by using the purified signals in Fig. 24. The final CM result is shown in Fig. 25.

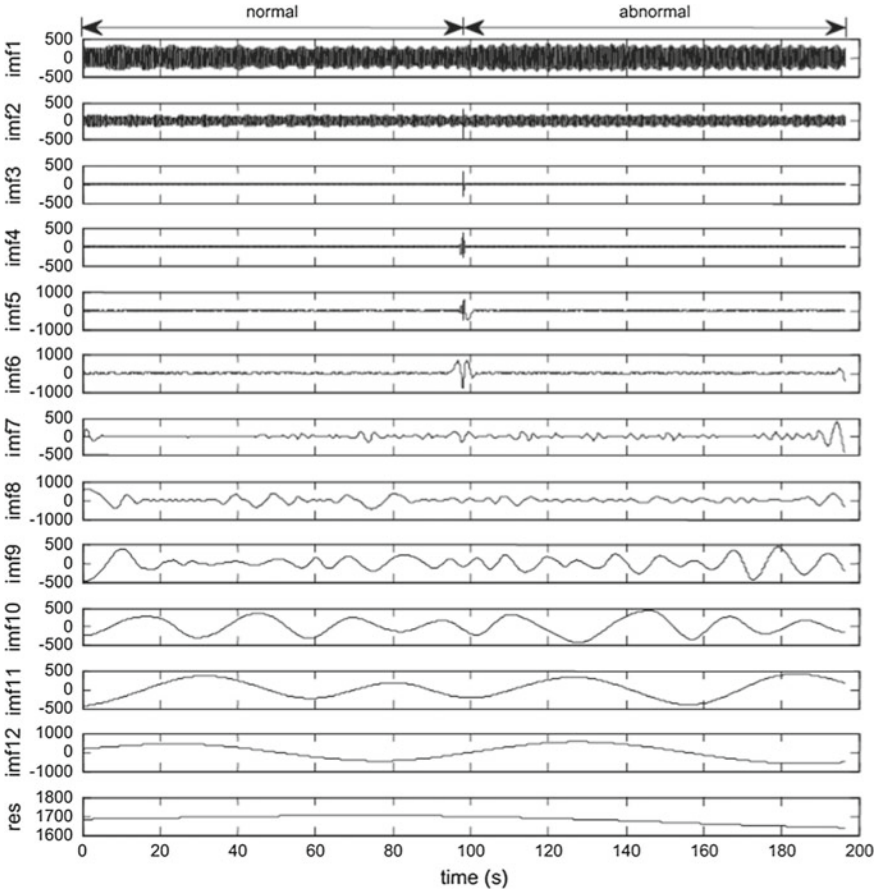
From Fig. 25, it is seen that in the presence of the 'mechanical unbalance fault', the calculated value of the CM criterion λ increases significantly, while the value of λ returns back normal level as soon as the fault is absent from the test rig. This demonstrates that the 'mechanical unbalance fault' has been successfully detected with the aid of the BEMD.

Based on the aforementioned two experiments, it can be concluded that

(1) The BEMD inherits all advantages of the EMD. The resultant zero-mean IMFs completely overcome or significantly mitigate the negative influences of wind turbine varying operational and loading conditions on its CM signals. Thus, the



(a) Electrical power signal



(b) The EMD results

Fig. 22 Detect mechanical unbalance fault by the approach of the EMD

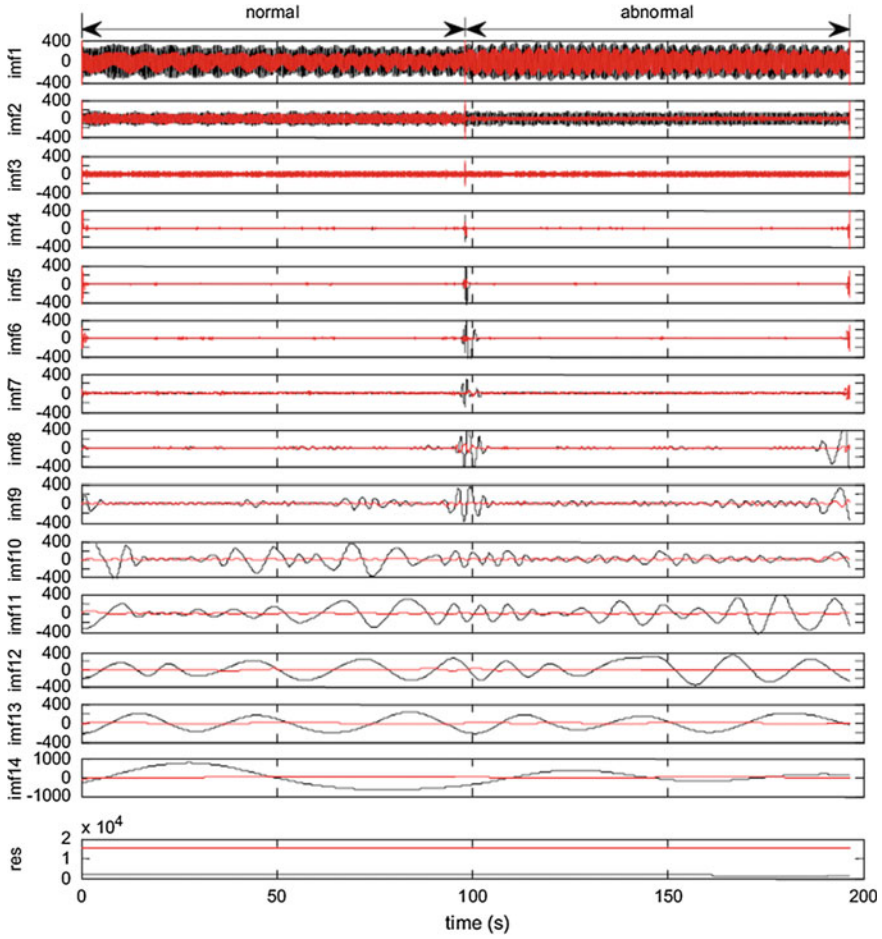


Fig. 23 The BEMD results for the signals in Fig. 15 [15]

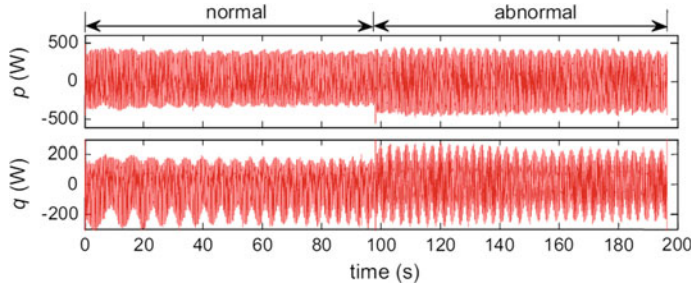


Fig. 24 The purification results based on the signal decomposition results in Fig. 17 [15]

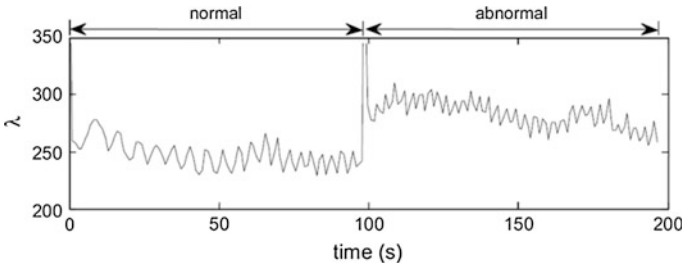


Fig. 25 CM result when mechanical unbalance fault occurs

application of the BEMD has potential to lead to a more reliable wind turbine CM conclusion;

(2) The BEMD-based CM criterion depicted by Eq. (15) is effective in detecting both the electrical and mechanical faults occurring in wind turbine drive train.

5 Concluding Remarks

From the work depicted above, the following concluding remarks can be made.

- The EMD enables a 'real-life' representation of the CM signal of interest attributed to its locally adaptive capacity that enable it can exactly match the signal features. The vivid and correct representation of the signal can significantly benefit machine CM, particularly in feature extraction and fault diagnosis. However, the EMD was designed only to process one-dimensional signals. Thus, it is unable to deal with information fusion issues in machine CM;
- The BEMD inherits all advantages of the EMD and moreover possesses an additional unique capacity of information fusion, attributed to which the BEMD effectively preserves the phase information of the signals in decomposition process. Thus, in contrast to the traditional EMD and the other TFA methods, the BEMD makes it more realistic than ever before to conduct reliable machine CM by the approach of information integration and/or fusion;
- As the resultant IMFs are zero-mean signals, both the EMD and BEMD overcome the negative influences of varying operational conditions on machine CM signals. This provides a new clue to develop operational-condition-independent CM techniques, which are very necessary and important for conducting the CM of variable speed machines, like wind turbines, helicopters and so on;
- Since the BEMD treats the signals collected from two transducers as one complex-valued signal, the BEMD is more efficient CM technique in computation than its counterpart EMD.

References

1. Wang, W., Mcfadden, P.D., "Application of wavelets to gearbox vibration signals for fault detection", Journal of Sound and Vibration 1996, 192:927–939.
2. Newland, D.E., "Ridge and phase identification in the frequency analysis of transient signals by harmonic wavelets", Transactions of ASME Journal of Vibration and Acoustics 1999, 121:149–155.
3. Tse, P., Yang, W.X., Tam, H.Y., "Machine fault diagnosis through an effective exact wavelet analysis", Journal of Sound and Vibration 2004, 277, (4–5):1005–1024.
4. Peng, Z., Chu, F., Tse, P., "Detection of the rubbing caused impacts for rotor-stator fault diagnosis using reassigned scalogram", Mechanical Systems and Signal Processing 2005, 19 (2): 391–409.
5. Yang, W.X., "A natural way for improving the accuracy of the continuous wavelet transforms", Journal of Sound and Vibration 2007, 306 (3–5):928–939.
6. Chen, X., Xiang, J., Li, B., He, Z., "A study of multiscale wavelet-based elements for adaptive finite element analysis", Advances in Engineering Software 2010, 41(2): 196–205.
7. Yan, R.Q., Gao, R., Chen, X.F., "Wavelets for fault diagnosis of rotary machines: A review with applications", Signal Processing 2014, 96:1–15.
8. Peng, Z., Tse, P.W., Chu, F.L., "A comparison study of improved Hilbert-Huang transform and wavelet transform: Application to fault diagnosis for rolling bearing", Mechanical Systems and Signal Processing 2005, 19:974–988.
9. Yang, W.X., "Interpretation of mechanical signals using an improved Hilbert-huang transform", Mechanical Systems and Signal Processing 2008, 22(5):1061–1071.
10. Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shin, H.H., Zheng, Q, Yen, N.C., Tung, C.C., Liu, H.H., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", Proceedings of Royal Society London, Part A 1998, 454:903–995.
11. Huang, N.E., Shen, Z., Long, S.R., "A new view of nonlinear water waves: the Hilbert spectrum", Annual Review of Fluid Mechanics 1999, 31:417–457.
12. Tanaka, T., Mandic, D.P., "Complex empirical mode decomposition", IEEE Signal Processing Letters 2007, 14 (2):101–104.
13. Rilling, G., Flandrin, P., Goncolves, P., Lilly, J.M., "Bivariate empirical mode decomposition", IEEE Signal Processing Letters 2007, 14 (12):936–939.
14. Yang, W.X., Tavner, P.J., "Empirical mode decomposition, an adaptive approach for interpreting shaft vibratory signals of large rotating machinery", Journal of Sound and Vibration 2009, 321(3–5):1144–1170.
15. Yang, W.X., Court, R., Tavner, P.J., Crabtree, C.J., "Bivariate empirical mode decomposition and its contribution to wind turbine condition monitoring", Journal of Sound and Vibration 2011, 330:3766–3782.
16. Yang, W.X., Tavner, P.J., Crabtree, C., Feng, Y., Qiu, Y., "Wind turbine condition monitoring: Technical and commercial challenges", Wind Energy 2014, 17 (5):673–693.
17. Zidani, F., Benbouzid, M.E.H., Diallo, D., Nait-Said, M.S., "Induction motor stator faults diagnosis by a current Concordia pattern-based fuzzy decision system", IEEE Transactions on Energy Conversion 2003, 18(4):469–475.
18. Yang, W.X., Tavner, P.J., Crabtree, C., Wilkinson, M., "Cost-effective condition monitoring for wind turbines", IEEE Transactions on Industrial Electronics 2010, 57(1):263–271.

Time-Frequency Demodulation Analysis Based on LMD and Its Applications

Yanxue Wang, Xuefeng Chen and Yanyang Zi

Abstract A time-frequency demodulation technique based on local mean decomposition (LMD) is proposed for rotating machine diagnosis. In addition, methods for boundary processing and for determining the step size of the moving average are presented to improve LMD algorithm. Instantaneous amplitude (IA) and instantaneous frequency (IF) of the signal can be achieved using the improved LMD method. A well-constructed description of the derived IAs and IFs is represented in the form of instantaneous time-frequency spectrum (ITFS), which preserves both the time and frequency information simultaneously. Results of three synthetic signals indicate that the proposed method is much better in extracting the comprehensive carrier and modulated components, compared with Hilbert-Huang transform and stationary wavelet transform. The validity of the technique is further demonstrated on the rotor system and a gearbox. The transient fluctuations of the IF and the impulsive signatures can be successfully identified in the ITFS. Moreover, it has been demonstrated that the proposed time-frequency demodulation technique is much more effective and sensitive than the other methods in detecting impulsive and modulated components.

1 Introduction

Modulated signals (AM/FM) widely exist in the vibration signal analysis and machinery fault diagnosis. Extracting the time-varying amplitude and frequency information from these signals are of great significance to determine the fault type and location. Hilbert transform and Teager energy operator (TEO) [1, 2] are

Y. Wang (✉)
Guilin University of Electronic and Technology, Guilin, People's Republic of China
e-mail: yan.xue.wang@gmail.com

X. Chen · Y. Zi
Xi'an Jiaotong University, Xi'an, People's Republic of China

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors,
Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_12

commonly used for demodulation analysis, but they are only suitable for processing mono-component signal. A variety of multi-component signal demodulation analysis techniques have been developed subsequently, such as multiband energy separation algorithm [3], periodic algebra separation, energy demodulation algorithm [4], iterated Hilbert transform [5, 6], complex shifted Morlet wavelet [7], EMD combined with TEO modulation [8, 9] and wavelet combined with TEO modulation [10]. Recently, a time-frequency manifold correlation matching for periodic fault identification in rotating machines was developed in [11].

Local mean decomposition (LMD) is an iterative decomposition method which was developed by Smith [12]. LMD provides a new idea to compute the instantaneous frequency (IF) and instantaneous amplitude (IA). However, there are several issues in the practical applications of the LMD. Thus, a new method of boundary process and a strategy for determining the step size of moving average (MA) are presented in this research, which improve the LMD method. Based on the improved LMD, a novel time-frequency demodulated method, which is independent of the HT, is proposed in this paper.

The rest of this paper is organized as follows. The instantaneous features of a signal are first introduced in Sect. 2. The theory of LMD is briefly introduced in Sect. 3. Section 4 discusses some key issues of LMD, such as boundary processing, determination of the step size of the moving average and the construction of instantaneous time-frequency spectrum. The validity of the LMD in the demodulation analysis is verified in Sect. 5 using simulated signals. Several practical applications of time-frequency demodulation are presented in Sect. 6. Conclusions are finally given in Sect. 7.

2 Instantaneous Features of a Signal

It's necessary to introduce the concept of IA and IF corresponding to AM/FM. For a mono-component signal

$$x(t) = a\cos(2\pi f \cdot t + \varphi_0) \tag{1}$$

Signal $x(t)$ is defined by three parameters, amplitude a, frequency f, and initial phase φ_0. For convenience, circular frequency ω ($\omega = 2\pi f$) is considered instead of f. Generally speaking, the amplitude and frequency are always the function of time, thus $x(t)$ can be written

$$x(t) = a(t)\cos[2\pi f(t) \cdot t + \varphi_0] \tag{2}$$

Because of the time-varying characteristic, $a(t)$ and $f(t)$ are named as IA and IF of signal $x(t)$, respectively. When $f(t)$ is changed with a nonzero rate of $\Delta f(t)$, that is, $f(t) = f_0 + \Delta f(t)$, and instantaneous phase (IP) can be written below

$$\varphi(t) = 2\pi f_0 \cdot t + 2\pi \int_0^t \Delta f(\varsigma) d\varsigma + \varphi_0 \tag{3}$$

in which $\varphi(t)$ is an IP. Thus, IF can be seen as first-order derivative of $\varphi(t)$. Moreover, IA can be obtained via the average of signal local instantaneous period amplitude.

$$a(t) = \sqrt{\frac{2}{T(t)} \int_{t-T(t)}^{t} x^2(\zeta) d\zeta} \tag{4}$$

where $T(t) = 1/f(t)$. Practically, signal demodulation with LMD method is in nature based on this idea, which can be seen in the following sections.

Indeed, the IP is also able to obtain by applying another solution of the analytic signal. Equation 2 can be abbreviated as $x(t) = a(t)\cos[\varphi(t)]$, for a non-analytic signal $x(t)$. There are many combination form $[a(t), \varphi(t)]$ which can generate $x(t)$. It's possible to get an analytic signal by using the original signal and its conjugation signal.

$$x^+(t) = x(t) + j\tilde{x}(t) \tag{5}$$

The conjugation signal $\tilde{x}(t) = \int_{-\infty}^{\infty} \frac{x(\tau)}{\pi(t-\tau)} d\tau$ is the Hilbert transform of $x(t)$. Therefore, the analytic signal $x^+(t)$ can be represented as $x^+(t) = a^+(t)e^{i\varphi^+(t)}$ via the unique pair of normalized amplitude and phase $[a^+(t), \varphi^+(t)]$ [13], in which $a^+(t)$ and $\varphi^+(t)$ are obtained respectively

$$a^+(t) = \sqrt{x(t) + \tilde{x}(t)}, \quad \varphi^+(t) = \arctan\frac{\tilde{x}(t)}{x(t)} \tag{6}$$

Moreover, IF is denoted as the derivative of phase of an analytic signal

$$f(t) = \frac{1}{2\pi}\frac{d\varphi^+(t)}{dt} = \frac{1}{2\pi}\frac{x(t)\dot{\tilde{x}}(t) - \dot{x}(t)\tilde{x}(t)}{x^2(t) + \tilde{x}^2(t)} \tag{7}$$

It should be noted that the approach for solving IA and IF mentioned above is widely used in EMD technology.

3 A Brief Introduction of LMD

LMD is an adaptive signal decomposition method and is first successfully applied to electroencephalogram signal [12]. LMD can decompose a complicate multi-component signal into a set of product functions (PFs). Each PF component is a product of a pure FM signal and an envelope signal. IF of each PF can be directly deduced from the pure FM signal with clear physical meanings, while the envelope signal is the IA. For a given signal, the process of LMD is written as follows,

(i) Determine all local maxima $n_{ij}(k_l), k_l = k_1, k_2, \ldots, k_M$ of original signal, k_l is the index of a maximum, M is the number of maxima, the subscript i indicates the ith PF and the subscript j represents the cycle number. So the local mean and local magnitude is written below

$$m_{ij}(t) = \frac{n_{ij}(k_l) + n_{ij}(k_{i+1})}{2}, \quad k_l = k_1, \ldots, k_{M-1}, t \in [k_l, k_{l+1}) \tag{8}$$

$$a_{ij}(t) = \frac{\left|n_{ij}(k_l) - n_{ij}(k_{l+1})\right|}{2}, \quad k_l = k_1, \ldots, k_{M-1}, t \in [k_l, k_{l+1}) \tag{9}$$

(ii) The calculated local mean $m_{ij}(t)$ and local magnitude $a_{ij}(t)$ gernerated by Eqs. (8) and (9) are shown in Fig. 1. Then, the local mean $\tilde{m}_{ij}(t)$ and magnitude $\tilde{a}_{ij}(t)$ are smoothed by using the moving average (MA).

(iii) The first smoothed local mean function $\tilde{m}_{11}(t)$ is separated from original signal $x(t)$, so we can obtain

$$h_{11}(t) = x(t) - \tilde{m}_{11}(t) \tag{10}$$

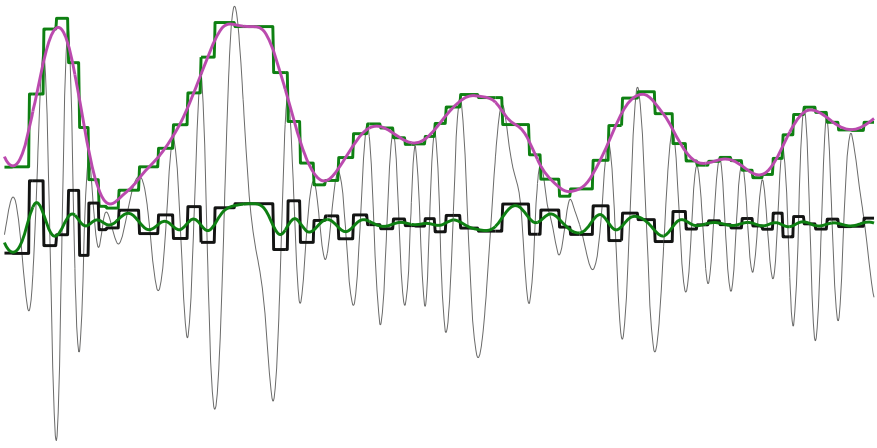


Fig. 1 The results of local mean and local magnitude

The FM component $s_{11}(t)$ is achieved using $h_{11}(t)$ divided by $\tilde{a}_{ij}(t)$,

$$s_{11}(t) = \frac{h_{11}(t)}{\tilde{a}_{11}(t)} \tag{11}$$

It is expected that $s_{11}(t)$ is a pure FM signal which oscillates in the interval of [–1, 1]. Thus, it is necessary to repeatedly perform the above process (i) $\sim$ (iii) until the expected FM signal $s_{1r_1}(t)$ is obtained. If r_1 represents iterative times corresponding to the first pure FM signal.

The whole iterative process can be represent by,

$$\begin{cases} h_{11}(t) = x(t) - \tilde{m}_{11}(t) \\ h_{12}(t) = s_{11}(t) - \tilde{m}_{12}(t) \\ \vdots \\ h_{1r_1}(t) = s_{1(r_1-1)}(t) - \tilde{m}_{1r_1}(t) \end{cases} \tag{12}$$

$$\begin{cases} s_{11}(t) = \frac{h_{11}(t)}{\tilde{a}_{11}(t)} \\ s_{12}(t) = \frac{h_{12}(t)}{\tilde{a}_{12}(t)} \\ \vdots \\ s_{1r_1}(t) = \frac{h_{1r_1}(t)}{\tilde{a}_{1r_1}(t)} \end{cases} \tag{13}$$

The stopping criterion for the iteration is

$$\lim_{r_1 \to \infty} a_{1r_1}(t) = 1 \tag{14}$$

(iv) Multiply all the smoothed envelop functions generated in the above iterative process, then the first IA function $a_1(t)$ is expressed as

$$a_1(t) = \tilde{a}_{11}(t)(t)\tilde{a}_{12}(t)(t) \ldots \tilde{a}_{1r_1}(t) \tag{15}$$

and its corresponding IP can be obtained by the first pure FM function using

$$\varphi_1(t) = \arccos(s_{1r_1}(t)) \tag{16}$$

By the derivation of IP, we can get the IF,

$$f_1(t) = \frac{f_s \cdot d\varphi_1(t)}{2\pi \cdot dt} \tag{17}$$

Multiply the first IA $a_1(t)$ with the pure FM signal $s_{1r_1}(t)$, the first PF component of original signal can be derived

$$PF_1(t) = a_1(t)s_{1r_1}(t) \tag{18}$$

An example of the IA, FM and PF achieved with Eq. 18 are shown in Fig. 2.

(v) Separate the first PF component $PF_1(t)$ from the original signal, we get the signal $u_1(t)$. Considering this signal as a new original signal, then repeat the above steps (i) $\sim$ (iv) p times until $u_p(t)$ becomes a monotone series

$$\begin{cases} u_1(t) = x(t) - PF_1(t) \\ u_2(t) = u_1(t) - PF_2(t) \\ \vdots \\ u_p(t) = u_{p-1}(t) - PF_p(t) \end{cases} \tag{19}$$

Thus, the original signal is decomposed into a set of PF components and a residual,

$$x(t) = \sum_{i=1}^{p} PF_i(t) + u_p(t) \tag{20}$$

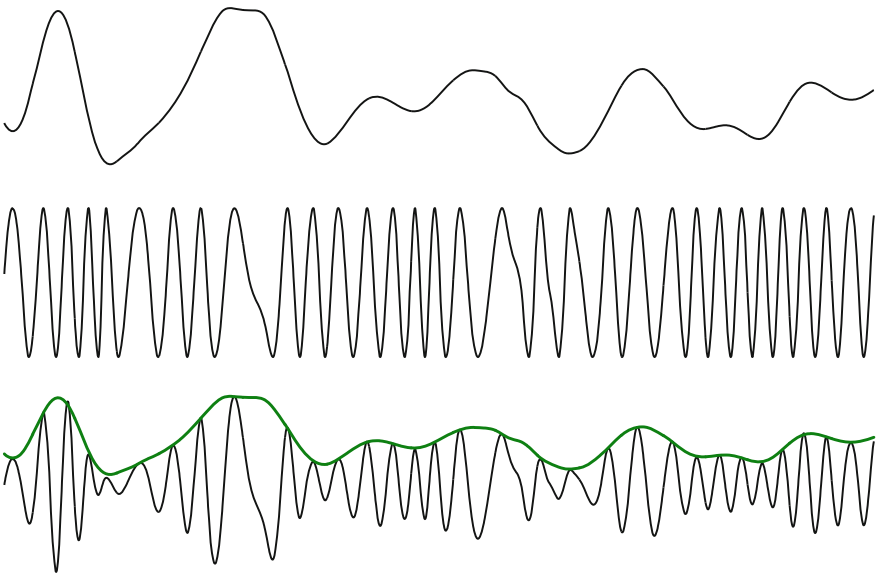


Fig. 2 The achieved three components: IA, pure FM and the correspongding PF

4 Some Key Issues of LMD

The theory of LMD algorithm has been discussed in the last section. There are several issues that require further attention for effective application of LMD, for example, boundary processing, determination of the step size of moving average and construction of instantaneous time-frequency spectrum.

4.1 Boundary Processing

As is well known, if the upper and lower envelopes mentioned in EMD are not well constructed, the ends of the time series will oscillate; then the end infection will propagate inwards and corrupt the subsequent lower frequency IMFs [14]. LMD adopts Eqs. 8 and 9 to estimate the local mean and local magnitude, and the endpoint of the signal is not definitely the extrema, thus finding appropriate end condition methods are also necessary in LMD. However, the issue of end effect is not mentioned in [12]. According to the above theory of LMD, we develops two methods for processing boundary in this work which each has their advantages and disadvantages.

(a) ***Mirror symmetric extension method***

The mirror symmetric extension is an effective method for boundary processing, which is widely used in EMD, WT and other signal analysis techniques. Supposed the discrete signal is:

$$X = [X(1), X(2), \ldots, X(n)], T = [T(1), T(2), \ldots, T(n)] = [t_1, t_2, \ldots, t_n], \quad (21)$$

First, find maxima and minima of $X(t)$ with respect to sequence subscript (I_m, I_n), their ordinate values (T_m, T_n) and their abscissa values (U, V), denoted as:

$$I_m = [I_m(1), I_m(2), \ldots, I_m(M)] \quad (22)$$

$$I_n = [I_n(1), I_n(2), \cdots I_n(N)] \quad (23)$$

$$T_m(i) = t_{I_m(i)}, U(i) = x_{I_m(i)}, i = 1, \ldots, M \quad (24)$$

$$T_n(i) = t_{I_n(i)}, V(i) = x_{I_n(i)}, i = 1, \ldots, N \quad (25)$$

(i) **Extension of left edge**

For the time series X, as is shown in Fig. 3, the procedure that procedure of the mirror extension is done below. In the case of $I_m(1) < I_n(1)$ and $X(1) > V(1)$, maximum $I_m(1)$ is considered as the center of symmetric for extension of left edge,

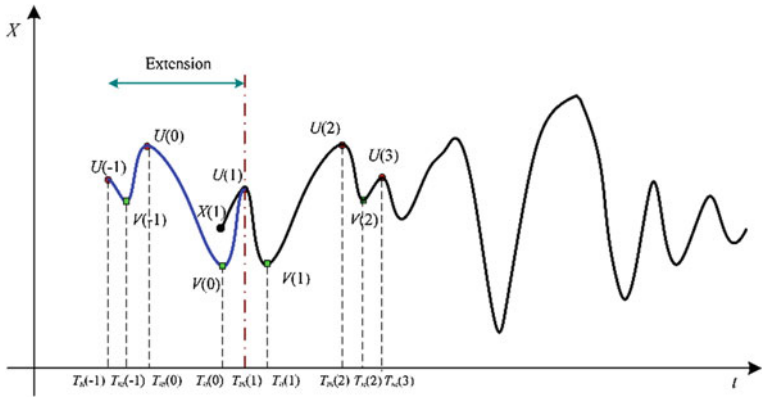


Fig. 3 In the case of $X(1) > V(1)$ the schematic figure for left edge extension of LMD

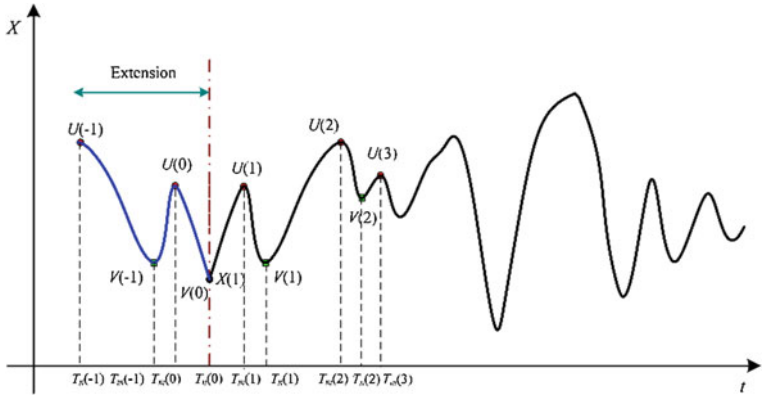


Fig. 4 In the case of $X(1) > V(1)$ the schematic figure for left edge extension of LMD

locations (T_m, T_n) and abscissa values of two successive minima and maxima are written as

$$T_m(0) = 2T_m(1) - T_m(2), \quad V(0) = V(2) \tag{26}$$

$$T_m(-1) = 2T_m(1) - T_m(3), \quad V(-1) = V(3) \tag{27}$$

$$T_n(0) = 2T_m(1) - T_n(1), \quad V(0) = V(1) \tag{28}$$

$$T_n(-1) = 2T_m(1) - T_n(2), \quad V(-1) = V(2) \tag{29}$$

Otherwise, if $\mathrm{X}(1) < V(1)$, as is shown in Fig. 4, left edge of time series will be regarded as the symmetric center for mirror extension. Thus, extended extrema locations (T_m, T_n) and their abscissa values (U, V) are as follows:

$$T_m(0) = 2t(1) - T_m(1), \quad U(0) = U(1) \tag{30}$$

$$T_m(-1) = 2t(1) - T_m(2), \quad U(-1) = U(2) \tag{31}$$

$$T_n(0) = t(1), \quad V(0) = X(1) \tag{32}$$

$$T_n(-1) = 2t(1) - T_n(1), \quad V(-1) = V(1) \tag{33}$$

In another case of $I_n(1) < I_m(1)$, the left endpoint may not be definitely the maximum or minimum. If $X(1) < V(1)$, *i. e*, $X(1)$ is the first minimum of series, so $I_n(1)$ will be regarded as the symmetric center for left mirror extension:

$$T_m(0) = 2t(1) - T_m(1), \quad U(0) = U(1) \tag{34}$$

$$T_m(-1) = 2t(1) - T_m(2), \quad U(-1) = U(2) \tag{35}$$

$$T_n(0) = t(1), \quad V(0) = X(1) \tag{36}$$

$$T_n(-1) = 2t(1) - T_n(1), \quad V(-1) = V(1) \tag{37}$$

Otherwise, if $\mathrm{X}(1) > U(1)$, $\mathrm{X}(1)$ should be regarded as the symmetric center for left extension using the following equations:

$$T_m(0) = t(1), \quad U(0) = X(1) \tag{38}$$

$$T_m(-1) = 2t(1) - T_m(1), \quad U(-1) = U(1) \tag{39}$$

$$T_n(0) = 2t(1) - T_n(1), \quad V(0) = V(1) \tag{40}$$

$$T_n(-1) = 2t(1) - T_n(2), \quad V(-1) = V(2) \tag{41}$$

(ii) **Extension of right edge**

For the right edge extension of the time series, it also needs to consider that whether the right edge is a maximum or a minimum, and the relationship between the right endpoint and the first extremum. The boundary processes in detail are similar to (**i**), thus it is not necessary to cover these again.

(b) ***Endpoint extrema prediction***

Symmetric extension is a common technique in restraining the end effect of LMD. However, it has some disadvantage, for example increasing the length of data and the time-consumption. Therefore, a new boundary processing method is proposed in [15], which uses the mean of extrema on the endings to obtain the controlled local magnitude and local mean. This boundary processing method could keep boundary from divergence and does not add additional computation. However, if insufficient data points both at the beginning and at the end of finite-duration signals

are used, the performance of this method is not better than that of mirror symmetric extension.

$$m_{ij}(t) = \frac{\left|n_{ij}(k_1) + 2n_{ij}(k_2) + n_{ij}(k_3)\right|}{4}, \quad t \in [k_0, k_1) \tag{42}$$

$$a_{ij}(t) = \frac{\left|n_{ij}(k_M) - n_{ij}(k_{M-1})\right|}{4} + \frac{\left|n_{ij}(k_{M-1}) - n_{ij}(k_{M-2})\right|}{4}, \quad t \in [k_0, k_1) \tag{43}$$

$$m_{ij}(t) = \frac{\left|n_{ij}(k_{M-2}) + 2n_{ij}(k_{M-1}) + n_{ij}(k_M)\right|}{4}, \quad t \in [k_M, k_{M+1}) \tag{44}$$

$$a_{ij}(t) = \frac{\left|n_{ij}(k_M) - n_{ij}(k_{M-1})\right|}{4} + \frac{\left|n_{ij}(k_{M-1}) - n_{ij}(k_{M-2})\right|}{4}, \quad t \in [k_M, k_{M+1}) \tag{45}$$

Actually, there are several methods in handling end-effect, but it is very difficult for a technique (including the methods proposed in this paper) to be always suitable for all the cases. Therefore, for the following applications of LMD in this paper, we will firstly attempt to adopt extreme prediction boundary processing, if necessary the mirror symmetric extension technique will be further considered.

4.2 Determination of the Step Size of MA

LMD method utilizes MA to generate smooth local mean and local magnitude. If the step size of MA is not appropriately selected, it can severely influence the decomposed PFs, IAs and IFs. Generally speaking, the step size of MA is set to one-third of the longest local mean, which is verified to be effective for the slow-varying filtered EEG data in [15]. When LMD method is applied to analyze the impact-type transient vibration signals, such as signals induced by local defect of bearing outer-race or inner-race, the impulsive features may be gradually polished with the increase of step size of MA, as is illustrated in Fig. 5. In fact, the

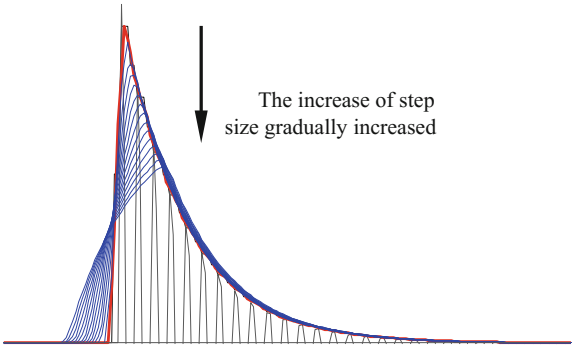


Fig. 5 The effects of step size of MA on impulsive signature

influence of step size used in MA is similar to that of wavelet function adopted in wavelet decomposition. If the selected wavelet function does not match the characteristic embedded in the analyzed signal, the expected features will not be extracted. However, EMD also suffers from a similar issue. Since many spline functions can be employed in the EMD technique, such as different order polynomial spline function, B-spline function and Hermite-spline function. Nevertheless, cubic spline function widely adopted in EMD is still chosen empirically.

In view of the above problems, a new strategy for the determination of the step size of MA is proposed in [15]. Since the extrema can be determined with Eq. 12 which is associated with $h_{i(j-1)}(t)$, so

$$\Delta(k_l) = k_l - k_{l-1}, l = 1, 2, \ldots, M+1 \tag{46}$$

$$ML = \max\Delta(k_l) \tag{47}$$

$$w = ML/R \tag{48}$$

where w is the step size of MA, and R is a constant. When LMD is used to analyze slow-varying signals, the results may be similar for different R (such as $R = 3$ or $R = 5$). While transient signal is analyzed, $R = 5$ or bigger value should be selected. Moreover, it should be noted that when R is greater, on the one hand it results in better extracting impulsive feature due to the smaller step of MA, on the other hand it increases the burden of processing. R is set to 3 in the following sections, and the step-size is determined using Eqs. 46–48.

4.3 Instantaneous Time-Frequency Spectrum

The time-frequency analysis technique can provide information of signal both in time domain and frequency domain simultaneously, therefore it has found a wider range of applications. Instantaneous time-frequency spectrum (ITFS) of LMD is based on the achieved IFs and IAs, which comprehensively reflects the changes of frequency and amplitude over time [15]. The IA of LMD can be obtained by Eq. 17, and IF is derived using Eq. 17, i.e., the first derivative of IP. Similar to the Hilbert-Huang spectrum (HHS) [14, 16], ITFS does not involve the concept of time- and frequency-resolution but IF. In order to calculate IF, the discrete IP function $\varphi_j(t)$ has been deconvoluted and the accurately derivation of IP is written as its numerical derivation method [17],

$$\dot{\varphi}_j[n] = \frac{1}{12}\left(\varphi_j[n-2] - 8\varphi_j[n-1] + 8\varphi_j[n+1] - \varphi_j[n+2]\right) \tag{49}$$

On the two endpoints of time series, we can adopt the derivation on the boundary [17], for example on the left ledge

$$\dot{\varphi}_j[n] = \frac{1}{12}\left(-25\varphi_j[n] + 48\varphi_j[n+1] - 36\varphi_j[n+2] + 16\varphi_j[n+3] - 3\varphi_j[n+4]\right) \tag{50}$$

and similarly, on the right ledge

$$\dot{\varphi}_j[n] = -\frac{1}{12}\left(-25\varphi_j[n] + 48\varphi_j[n-1] - 36\varphi_j[n-2] + 16\varphi_j[n-3] - 3\varphi_j[n-4]\right) \tag{51}$$

Moreover, smooth window average technology mentioned in [18] can be utilized to calculate IF via Kaiser window function is written as

$$h[n] = \frac{1.5N}{N^2 - 1}\left(1 - \left(\frac{n - (0.5N - 1)}{0.5N}\right)^2\right) \tag{52}$$

$$\dot{\varphi}_j[n] = \sum_{n=1}^{N-1} h[n]\left(\varphi_j[n+1] - \varphi_j[n]\right) \tag{53}$$

Different from EMD method, the HHS can be structured by Hilbert transform based on the generated IMFs. However, IA and IF have been directly obtained, ITFS of LMD is constructed independent of Hilbert transform along with decomposition. The derived ITFS for the time series X can be expressed as follows

$$ITFS_X(t,f) = \sum_{j=1}^{J} a_j\left(t, f_j(t)\right) = \sum_{j=1}^{J}\sum_{i} a_j(t)\delta\left(f - f_{j_i}(t)\right) \tag{54}$$

ITFS comprehensively reflects the extracted information. Therefore, its applications in demodulation of simulated and practical signals are investigated in the following sections.

5 Time-Frequency Demodulation Analysis for the Simulated Signals

The AM and FM signals are simulated and adopted here. It is very difficult to distinguish them in frequency, for they have almost similar spectra. Effectiveness of ITFS in time-frequency demodulation is conducted in this section, compared with

Hilbert-Huang Spectrum (HHS) [14] and wavelet projection Hilbert spectrum (WPHS) [17].

5.1 AM Signal Demodulation

An AM signal $s_1(t)$ is represented as follows

$$s_1(t) = cos(2\pi \cdot 400 \cdot t) * (1 + 0.5\cos(2\pi \cdot 25 \cdot t)), \quad t \in [0, 0.512] \tag{55}$$

Figure 6 illustrates its time-domain waveform and spectrum. It can be seen in the spectrum there is a dominant 400 Hz frequency associated with 25 Hz sideband. Figure 7a shows the decomposed PF components, where one can find the boundary is done well using the proposed technique. The ITFS of LMD is shown in Fig. 7b where it can be clearly seen that LMD well extracts the carrier and modulated components of the designated signal. In addition, the derived IF of the carrier (400 Hz frequency component) is illustrated as a horizontal line with periodically changed magnitude (25 Hz modulator) which indicates it is an AM signal. The obtained HHS and WPHS as shown in Fig. 7c, d, respectively. We can find the HHS is well to extract the 25 Hz modulator component, but the carrier is not obvious. Though WPHT can better extract carrier component, but modulator information is missing.

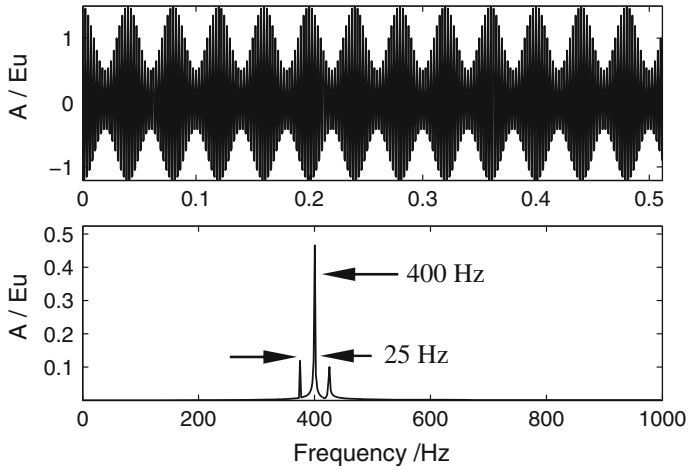


Fig. 6 The simulated AM signal and its spectrum

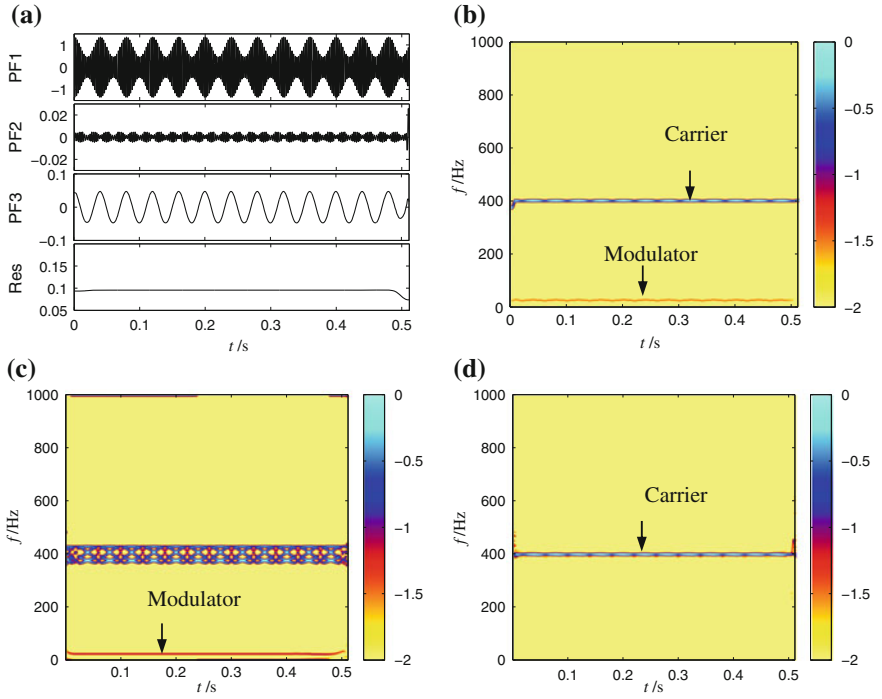


Fig. 7 Time-frequency demodulation for simulated AM signal **a** temporal signal, **b** its ITFS, **c** HHS and **d** WPHS

5.2 *FM Signal Demodulation*

The $s_2(t)$ is a FM signal and is expressed as

$$s_2(t) = \cos(2\pi \cdot 400 \cdot t + 0.5\cos(2\pi \cdot 25 \cdot t)), \quad t \in [0, 0.512] \tag{56}$$

Figure 8 shows its time-domain waveform and spectrum. We can find the dominant frequency and sideband in Fig. 8 are the same with those shown in Fig. 6. Figure 9a, b show that PF components and ITFS achieved by LMD. In the ITFS, we can clearly find the 400 Hz carrier component and 25 Hz modulator component. Different from the result shown in Fig. 7b, both the carrier and modulator components fluctuate (i.e., frequency changes with time), which is actually the nature of an FM signal. Results of HHS and WPHS are shown in Fig. 9c, d, respectively. Nevertheless, information can be only partly identified from HHS and WPHS, which can be seen in Fig. 9c, d.

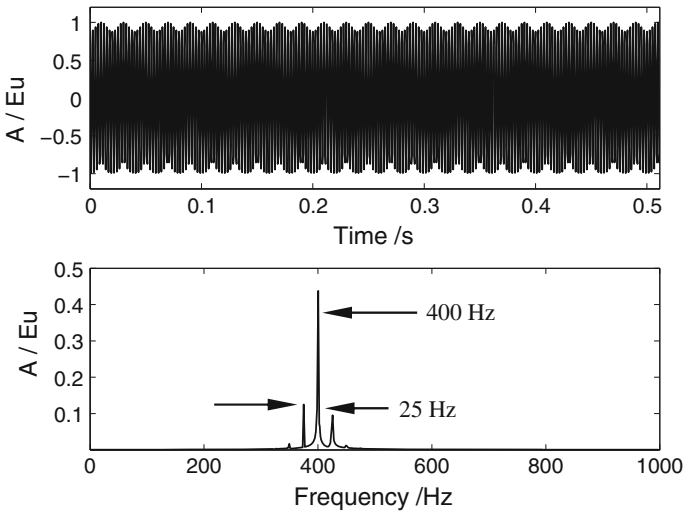


Fig. 8 The simulated FM signal and its spectrum

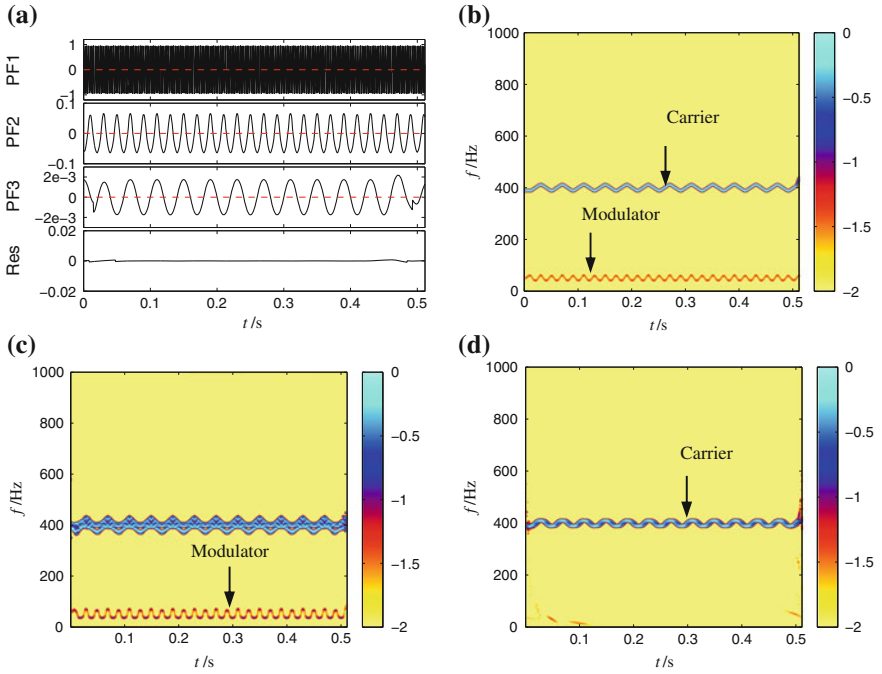


Fig. 9 Time-frequency demodulation for simulated FM signal **a** temporal signal, **b** its ITFS, **c** HHS and **d** WPHS

5.3 AM-FM Signal Demodulation

If $s_3(t)$ is a hybrid modulation signal:

$$s_3(t) = \cos(2\pi \cdot 400 \cdot t + 0.5\cos(2\pi \cdot 25 \cdot t)) \times (1 + 0.5\cos(2\pi \cdot 25 \cdot t)) \quad (57)$$

Time-domain waveform of the simulated signal and its spectrum are shown in Fig. 10. It can be seen its spectrum is very similar to above mentioned cases of AM and FM signals. Therefore, the modulation information of the signal is very difficult to be accurately identified from the spectrum of these three cases. At the same time, Fig. 11a, b show the achieved PF components and the corresponding ITFS. In the instantaneous time-frequency, it is easy to find that the carrier component locates at 400 Hz and its modulated frequency is 25 Hz (40 ms). As is mentioned above, the fluctuation of the modulator shows that there is a frequency modulation of the signal, while the alternation of color reflects amplitude modulation of the signal. Modulator component can also be seen around 25 Hz. As is shown in Fig. 11c, the 25 Hz modulator component can be seen in the result of HHS, but the carrier cannot be detected. The result of WPHS is illustrated in Fig. 11d, where the carrier information can be identified, but modulator cannot be found.

Through the simulation analysis, we can find that the LMD has a strong time-frequency demodulation ability compared with the HHS and WPHS, and LMD can simultaneously identify all the modulation information embedded in signals. Thus, time-frequency demodulation using LMD is very suitable for

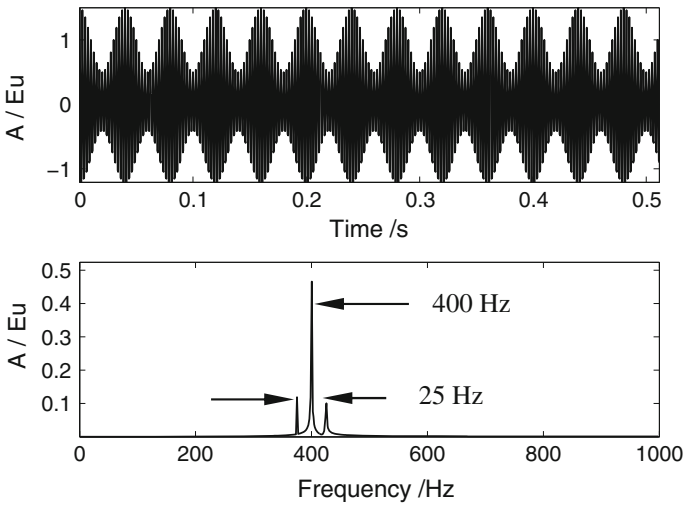


Fig. 10 The simulated AM-FM signal and its spectrum

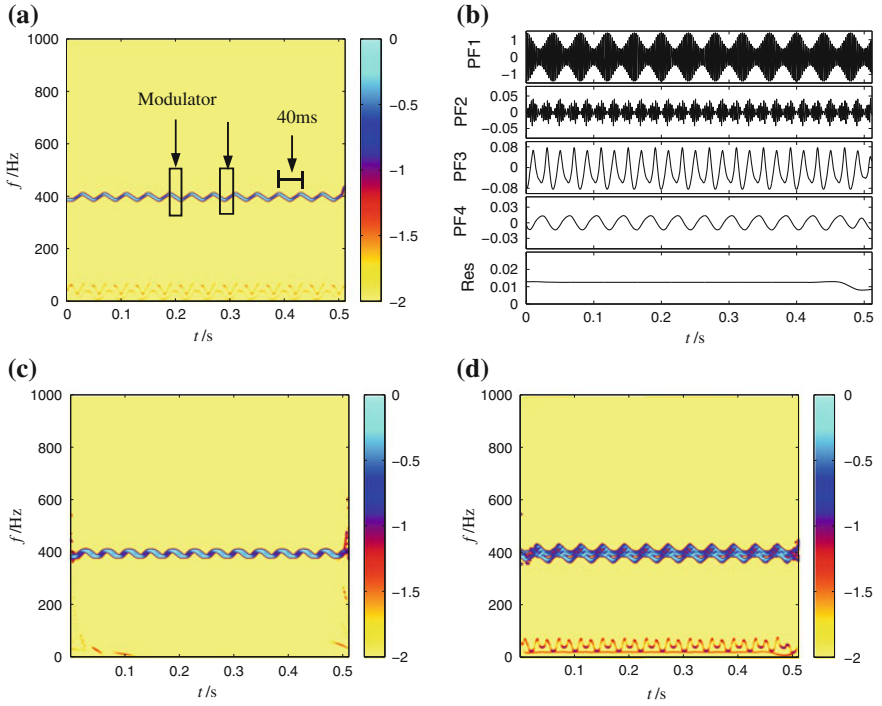


Fig. 11 **a** LMD decomposition of the simulated AM-FM signal and **b** Its Instantaneous Time-frequency, **c** HHS and **d** WPHS

processing mechanical non-stationary signals with multi-components. This characteristics will be further demonstrated using practical vibration signals acquired in a rotating machine.

6 Applications

The proposed time-frequency demodulation with LMD is utilized to detect the rub-impact fault in a rotor system and a practical rotating machine, as well as to identify damage in a gearbox.

6.1 Rub Fault Detection in a Rotor System

Rub occurs in rotating machinery with radial clearance between rotor and stator, such as bearing internal clearance, seal/packing or blade/case. Radial clearance

between rotor and stator in high-speed rotating machines such as generator sets, aeroengines, turbines and compressor is of great importance. If the radial clearance between the rotating rotor and the stator is smaller, the efficiency of these kinds of machines improves. The smaller the clearance of the modern rotating machinery, the more the possibility of rubbing occurs. If the rub induces the rotor's dynamic instability, a serious accident occurs. Therefore, it is important to detect rub by using signal processing techniques as early as possible. Rubs may cause impacts, sub-synchronous and super-synchronous vibrations of the shafts. Complicated nonlinear behavior is generally associated with a vibrating system of a faulty rotor and the vibration signals contain very complicated phenomenon including not only the periodic motion but also the chaotic motion. Rub signal was acquired by eddy current transducer on the Bently test-rig shown in Fig. 12. Sampling frequency was set to 2000 Hz and the number of sampling points was 1024. Moreover, friction rod was used to generate a slight local rub fault.

When the rotating speed is set to 2200 rpm, time-domain signal and its frequency spectrum are illustrated in Fig. 13a. Due to a slight rub fault, it is difficult to find the abnormal characteristics in time-domain waveform and its spectrum. Results of time-frequency demodulation of ITFS, HHS, WPHS are shown in Fig. 13b–d. We can find both ITFS and HHS can successfully detect FM components. Actually, the reason for this frequency fluctuation is caused by the friction between the friction rod and the rotor which will restrict the rotation of the rotor and decrease its rotating speed.

6.2 *Practical Rub-Impact Fault Diagnosis*

In this subsection, the proposed time-frequency demodulation method will be further applied in detecting rub-impact in a machine set named heavy oil catalytic cracking process. This machine set consists of a gas turbine, compressor, gearbox

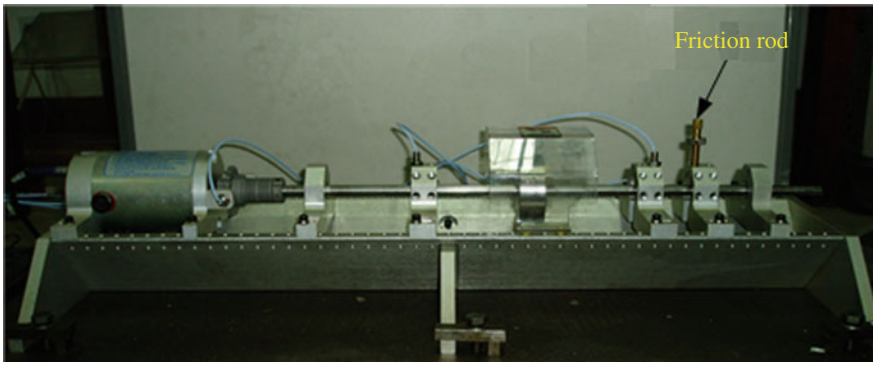


Fig. 12 The test-rig with rotor local rub fault

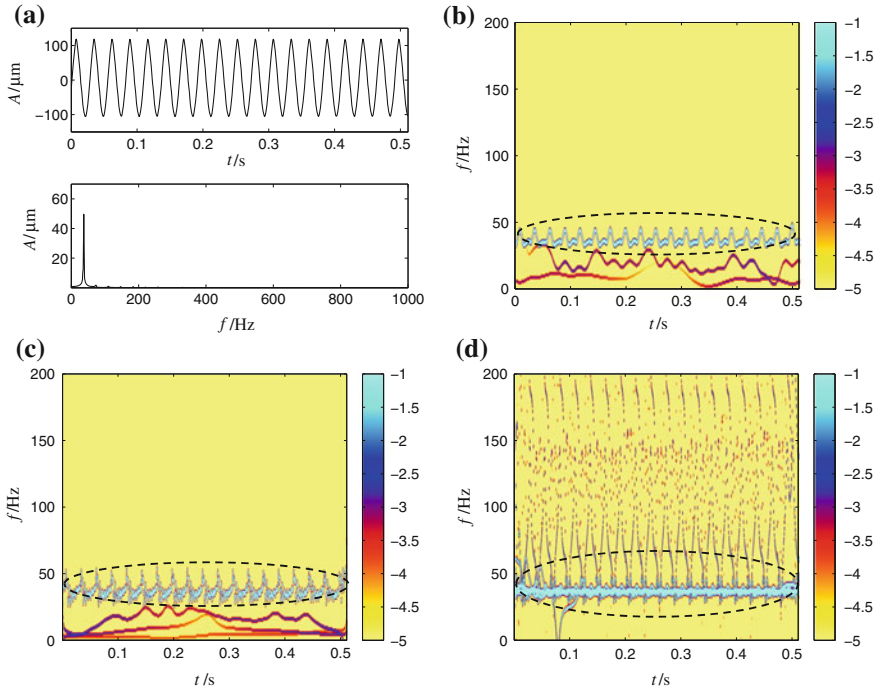


Fig. 13 Experimental verification **a** the rub-impact signal and its spectrum, **b** ITFS, **c** HHS and **d** WPHS

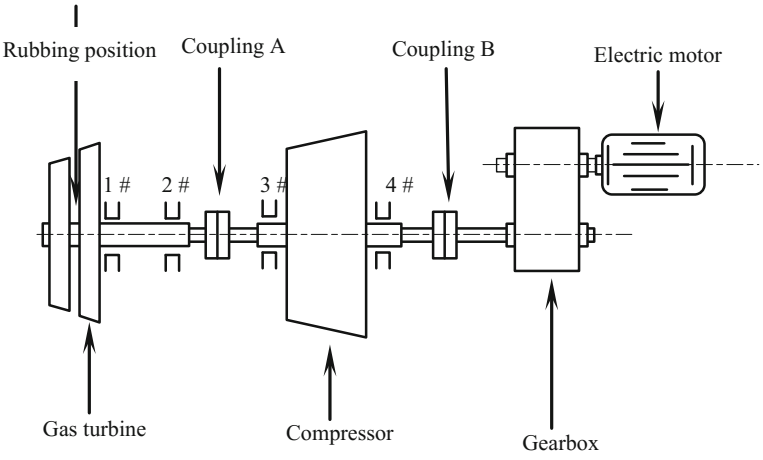


Fig. 14 The structural sketch of the machine set

and an electric motor. Figure 14 shows its structural sketch. The rotating speed of the gas turbine is 5745 r/min. Eddy current transducers are mounted in 2# bearing to pick up displacement vibration signals.

Figure 15a shows the measured vibration signal at a sampling rate of 2000 Hz. As can be seen in the temporal waveform and its FFT spectrum given in Fig. 15a, it is difficult to find the abnormal information except the fundamental harmonic component 1X and its second harmonic 2X. As is well known, when the rub-impact fault occurs under certain circumstances, especially at its early stage, the X/2 or X/3 etc. sub-harmonic components should be observed, besides the multiple harmonic components such as 2X, 3X etc. The achieved time-frequency distributions of LMD is presented in Fig. 16. To clearly show the details of the significant components in the ITFS, the range of the frequency axis is set to [0, 500] Hz. It can be seen that the proposed time-frequency demodulation technique can well detect the existence of the sub-harmonic component which oscillates around one third of 1X. In addition, the sudden change of the IF of the 1X can be identified from ITFS in Fig. 16 (marked by white rectangular), which actually reflects the effect of the opposite friction during operation. Consequently, the above results provide sufficient evidences to judge the existence of an early local rub-impact fault in this machine. Meanwhile, a new feature of rub-impact fault, the FM component is also identified around 1X component with the proposed method, whose modulation period is about 10.24 ms. Modulation frequency is 97.7 Hz (=1/10.24 ms), which equals to the fundamental harmonic component (=1X). These characteristics all indicate that the unit may have slight rub-impact fault. In addition, one more important information, when the frequency and X/3 times harmonic occurs down to the mutation, it will stimulate the characteristics that are similar to shocks (as is shown in the

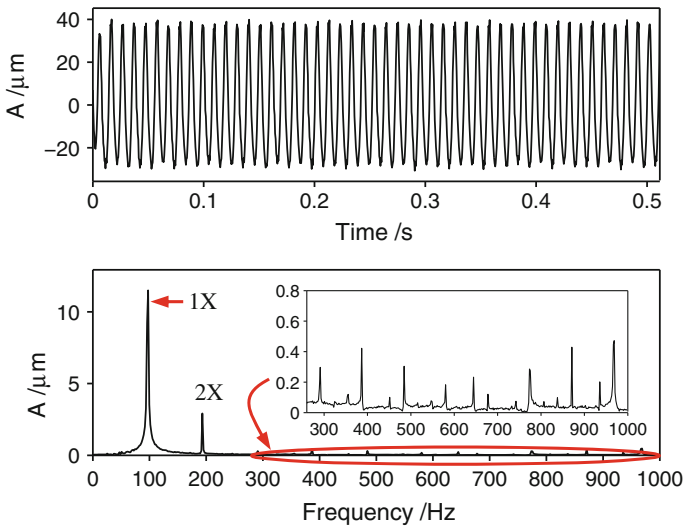


Fig. 15 The acquired vibration signal and its spectrum

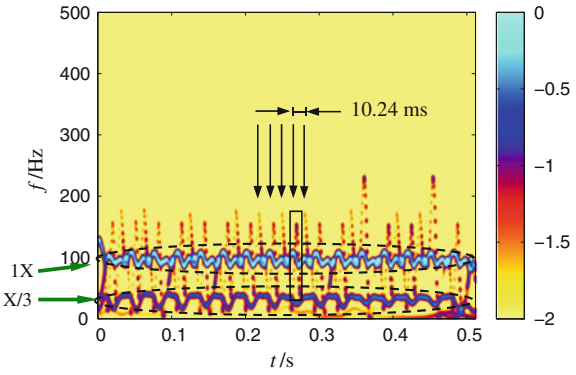


Fig. 16 The ITFS of the signal in Fig. 15a using the proposed time-frequency demodulation technique

rectangular frame). The reason is that when rub-impact occurs, the corresponding friction force is opposite to the rotating direction, thus the rotating frequency and the sub-harmonic frequency will decrease naturally.

Figure 17a, b show the results of HHS and WPHS, respectively. However, no distinct features can be observed in Fig. 17a, b except the 1X component. To sum up, the proposed time-frequency demodulation technique can successfully detect the instantaneous features of the signal and is much more powerful in detecting incipient rub-impact fault than the other two methods.

6.3 Demodulation Analysis of a Gearbox

Gearbox used in this subsection is a key equipment in the hot strip finishing rolling mill and it can directly affect the products as well as the long-term safety in operation. The main driven chain comprises a one-stage helical gearbox with a ratio of 2.9545 which usually works in a low rotating speed (its high speed shaft frequency is only about 2–5 Hz) and in a varying load conditions. Vibration signals were measured by velocity transducers attached to the outer casing of the gearbox. Signals in vertical direction are used in this work. The sampling frequency is set to 2560 and the number of sampling is 4096. The gear and pinion are all helical and their tooth numbers are 65, 22 respectively, and modulus is 30. Because the rolling mill process speed fluctuated in a range, the speed of high-speed axis was about 3–4 Hz.

Figure 18a shows a vibration signal and its spectrum which was acquired around the high-speed axis on the April 9, 2008. We can easy find the meshing frequency is 89.50 Hz in the spectrum, as well as its corresponding high-speed axis rotation frequency 4.068 Hz. As is well known, local fault gives impulse perturbations in the signal whose frequency is equivalent to gear shaft rotation. Thus, the proposed ITFS of the LMD is applied to demodulate the weak impulsive components in time-frequency domain. Figure 18b shows the ITFS of the temporal signal in

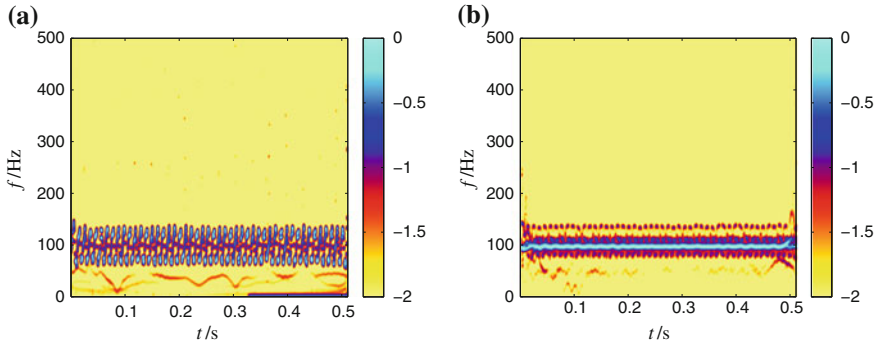


Fig. 17 **a** HHS and **b** WPHS of the signal in Fig. 15a

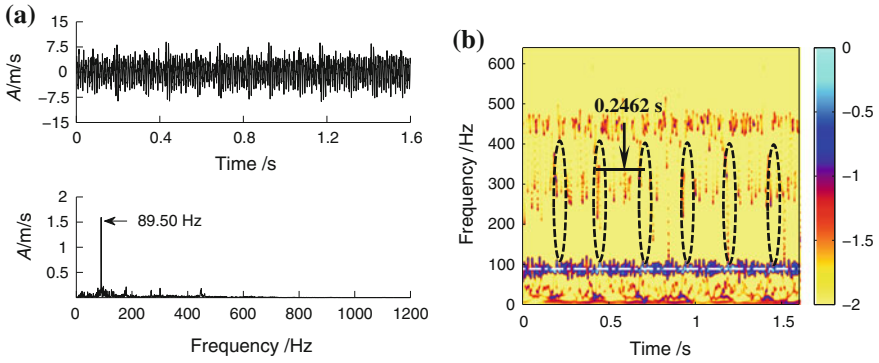


Fig. 18 **a** The gear box signal of rolling mill and its spectrum, **b** The ITFS Analysis of LMD In April 9, 2008

Fig. 18a, where meshing frequency (white dot line) and periodical impulsive components (circled with black dash lines) spaced by 0.2462 s (=4.068 Hz) can be found. The frequency of periodical impacts is the same with the rotating speed of pinion, which means the possible location of a local defect.

To further demonstrate the result, ITFS of another acquired signal is shown in Fig. 19. We can find meshing frequency (also white dot line) is changed to 100.6 Hz, as well as much clearer periodical impulsive features (circled with black dash lines) spaced by 0.2196 s whose frequency is 4.573 Hz (=100.6/22). The effectiveness of time-frequency demodulation of LMD is verified once more. We can draw a conclusion that high-speed axis gear has a local fault. Local scuffing on the pinion gear teeth has been observed in the following maintenance actions and then a new pinion was replaced as soon as possible. Figure 20 displays the damaged pinion. The ITFS of the vibration signal acquired after the pinion was replaced is shown in Fig. 21 where the periodical impulsive signatures cannot be observed.

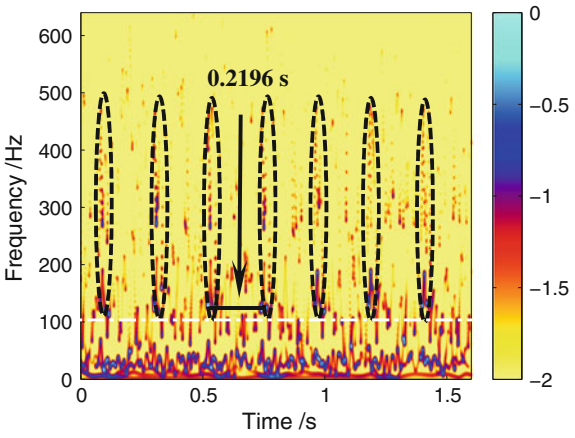


Fig. 19 ITFS demodulation analysis

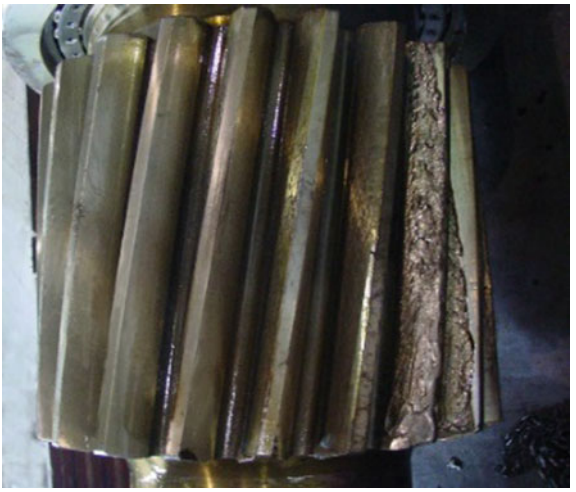


Fig. 20 The picture of the local scuffing damage

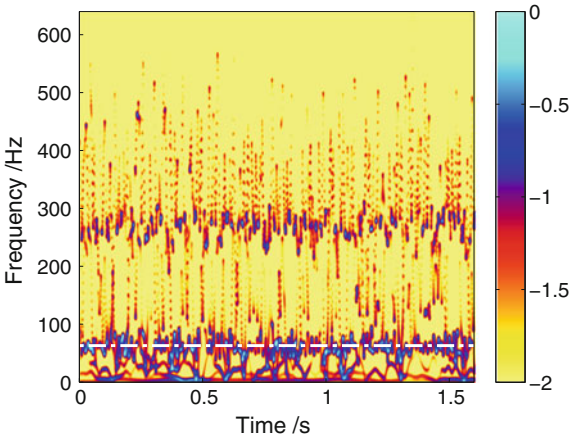


Fig. 21 ITFS of the vibration signal in the normal condition of mill

7 Conclusions

This work investigates the several key issues corresponding to LMD technique and its time-frequency demodulation analysis for fault detecting applications. We give two methods to resist end-effect and develops an adaptive strategy for determining the step size of MA used in LMD. Then, independent of Hilbert transform, ITFS is constructed for time-frequency demodulation. Simulated AM and FM signals are utilized to demonstrate the performance of time-frequency demodulation in comparison with HHS and WPHS. Results show time-frequency demodulation with LMD can comprehensively detect the modulator and carrier information. The proposed technique is then used in detecting rotor rubbing and gearbox local damage. These two practical applications further verify its effectiveness in complex multi-component signal demodulation analysis.

Acknowledgements The financial sponsorship from the project of National Natural Science Foundation of China (51475098 and 61463010) and Guangxi Natural Science Foundation (2016GXNSFFA380008) are gratefully acknowledged. It's also sponsored by Guangxi Experiment Center of Information Science (20130312) and Guangxi Key Laboratory of Manufacturing System and Advanced Manufacturing Technology (15-140-30-001Z).

References

1. Maragos P., Kaiser J.F., Quatieri T.F., "On amplitude and frequecy demodulation using energy operator", IEEE Transactions on Signal Processing, 1993, 41:1532–1550.
2. Potamianos A., Maragos P., A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation, Signal Processing, 1994, 37: 95–120.
3. Bovik A.C., Maragos P., Quatieri, T.F., "AM-FM energy detection and separation in noise using multiband energy operators", IEEE Transactions on Signal Processing, 1993, 41: 3245–3265.
4. Santhanam B., Maragos P., "Multicomponent AM-FM demodulation via periodicity-based algebraic separation and energy-based demodulation", IEEE Transactions on Communications, 2000, 48:473–490.
5. Gianfellci F., Biagetti G., Crippa P., Turchetti C., "Multicomponent AM-FM representations: an asymptotically exact approach," IEEE Trans. on Audio, Speech, and Language Processing, 2007, 15:823–837.
6. Qin Y., Qin S., Mao Y., "Research on iterated Hilbert transform and its application in mechanical fault diagnosis," Mechanical Systems and Signal Processing, 2008, 22(8): 1967–1980.
7. Nikolaou N.G., Antoniadis I.A., "Demodulation of vibration signals generated by defects in rolling element bearings using complex shifted Morlet wavelets," Mechanical Systems and Signal Processing, 2002, 16:677–694.
8. Cui S.., Li X.L., Ouyang G.X., Guan X.P., "Detection of epileptic spikes with empirical mode decomposition and nonlinear energy operator," *Advances in Neural Networks,* vol. 3498, ed. Berlin: Springer-Verlag Berlin, 2005, pp. 445–450.
9. Cheng J.S., Yu D. J., Yang Y., "The application of energy operator demodulation approach based on EMD in machinery fault diagnosis," Mechanical Systems and Signal Processing, 2007, 21:668–677.

10. Liu X. Q., “Hybrid wavelet packet-teager energy operator analysis and its application for gearbox fault diagnosis”, Chinese Journal of Mechanical Engineering, 2007, 20(6): 79–83.
11. He Q., Wang, X., “Time-frequency manifold correlation matching for periodic fault identification in rotating machines,” Journal of Sound and Vibration, 2013, 332:2611–2626.
12. Smith J.S., “The local mean decomposition and its application to EEG perception data,” Journal of the Royal Society Interface, 2005, 2(5):443–454.
13. Picinbono B., “On instantaneous amplitude and phase of signals,” IEEE Transactions on Signal Processing, 1997, 45:552–560.
14. Huang N.E., Shen Z., Long S. R., “The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis,” Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 1998, 454:903–995.
15. Wang Y., He Z., Zi Y., “A demodulation method based on local mean decomposition and its application in rub-impact fault diagnosis,” Measurement Science & Technology, 2009, 20, p. (10 pages).
16. Huang N.E., Shen Z., Long S. R., “A new view of nonlinear water waves: the Hilbert spectrum,” Annual Review of Fluid Mechanics, 1999, 31:417–457.
17. Olhede S., Walden A.T., “The Hilbert spectrum via wavelet projections,” Proceedings of the Royal Society of London Series A-Mathematical Physical and Engineering Sciences, 2004, 460:955–975.
18. Kay S., “Fast and accurate single frequency estimator,” IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989, 37:1987–1990.

On the Use of Stochastic Resonance in Mechanical Fault Signal Detection

X.F. Zhang, N.Q. Hu, L. Zhang, X.F. Wu, L. Hu and Z. Cheng

Abstract This chapter focuses on the application of stochastic resonance (SR) in mechanical fault signal detection. SR is a nonlinear effect that is now widely used in weak signal detection under heavy noise circumstances. In order to extract characteristic fault signal of the dynamic mechanical components, SR normalized scale transform is presented and a circuit module is designed based on parameter-tuning bistable SR. Weak signal detection based on stochastic resonance (SR) can hardly succeed when noise intensity exceeds the optimal value of SR. Therefore, a signal detection model based on combination effect of colored noise SR and parallel bistable SR array, which is called multi-scale bistable stochastic resonance array, has been constructed. Based on the enhancement effect of the constructed model and the normalized scale transformation, weak signal detection method has been proposed. The effectiveness of these methods are confirmed and replicated by numerical simulations. Applications of bearing fault diagnosis show the enhanced detecting effects of the proposed methods.

1 Introduction

Weak signal detection under heavy background noise is one of the focuses in various signal-processing fields. It is commonly concerned by scientists and engineers to detect or enhance weak target signal more expeditiously and precisely in noisy environment with certain restrictions. In the real-world systems, because characteristic signals of mechanical component early fault contain a little energy and are usually annoyed by heavy noise, it is a great challenge to reveal the characteristic signal. Effective characteristic signal detection approach is significant to the fault diagnosis of mechanical component, especially when the fault is in its

X.F. Zhang · N.Q. Hu (✉) · L. Zhang · X.F. Wu · L. Hu · Z. Cheng
National University of Defense Technology, Changsha, Hunan, People's Republic of China
e-mail: hnq_5744@163.com

R. Yan et al. (eds.), *Structural Health Monitoring*, Smart Sensors,
Measurement and Instrumentation 26, DOI 10.1007/978-3-319-56126-4_13

early stage. Prognosis of critical mechanical component mandates detecting the defect signatures as early as possible, so that the corresponding maintenance can be scheduled and the possible catastrophic accident or machine breakdown can be avoided. Consequently, detection of characteristic signal has become one of the key technologies of early fault diagnosis for mechanical component.

A traditional way of weak signal detection generally focuses on suppressing the noise to improve the output signal-to-noise ratio (SNR). Compared with conventional linear methods, the methods based on stochastic resonance (SR) are promising in the conditions of short data records and heavy noise. SR is a nonlinear effect that has been widely used in weak signal detection. Since proposed by Benzi et al. [1], stochastic resonance has been developing rapidly in signal processing, detection and estimation [2–9], especially in low SNR cases. It is essentially a statistical phenomenon resulting from an effect of noise on information transfer and signal processing that is observed in both man-made and naturally existing nonlinear systems. The counterintuitive SR phenomenon is caused by cooperation of signal (deterministic force) and noise (stochastic force) in a nonlinear system. In a certain nonlinear system, noise plays a constructive role, and energy flows from noise to signal. When noise or system parameters are tuned properly, the output SNR will reach a maximum.

However, the application of SR to practical problems has been restricted by the fact that the bistable system is only sensitive to low frequency and weak periodic signals. This can be explained in a formal way by adiabatic approximation and linear response theory [2]. In order to apply SR to high frequency signal detection, several system parameter tuning or noise intensity tuning methods have been proposed to make them more adaptive, such as normalized scale transform [10, 11], re-scaling frequency SR [12], frequency-shifted and re-scaling SR [13], adaptive step-changed [14], etc. In order to apply SR methods to characteristic vibration signal detection, two methods are discussed in this chapter: (1) normalized scale transform, a complete computation method for sampled vibration signal; (2) parameter-tuning SR circuit module for analog application. Basic theory of the two methods is parameter tuning SR. Simulations are made to validate the enhancing effect of the two methods.

The equivalence between noise tuning SR and parameter tuning SR in a typical bistable system with an additive white noise has been addressed in reference [6]. Only when the input noise intensity reaches the system resonance region, the system response is capable of following the input signal so that the output SNR is enhanced to conform to the nonlinear mechanism [5, 6]. In practice, tuning noise intensity is not always feasible. The intensities of signal and noise have been fixed for the collected raw signal in a practical engineering system. Using noise tuning, parameter tuning or array SR alone may not be a suitable option when the noise intensity exceeds the SR resonance region, which is often the case for digital signal processing and weak signal detection, such as the vibration signals collected from a gearbox for fault diagnosis and health state assessment. Besides tuning noise intensity and parameter, SR based signal processing could be improved for a better performance. The first approach is the cascaded bistable system [15],

which connects two or more bistable systems in series. The second one is the coupled or uncoupled parallel bistable array [16–19]. The third one is to make use of the characteristics of colour noise [20–22].

The SR effect can be driven not only by white noise but also by the band-limited noise alone, which indicates that it is possible to realize the SR by tuning the band limited noise. In addition, the array SR theory indicates that an array of bistable dynamical subsystems constructs a meaningful collective system for further improvement of output SNR [16–19]. In order to process the noisy signal that is beyond the SR resonance region, a summing SR array model called multi-scale bistable array (MSBA) is constructed, which consists of several bistable units. This parallel bistable array model also combines normalized scale transform with inherent SR effect driven by colour noise. At first, the processed signal is decomposed into some different scale signals by wavelet transform. Each unit is subject to different scale noise, which plays the role of inner noise of the array. The scale signal containing the target signal is processed as the noisy input signal of the array. By summing the output signal from each unit, we can obtain a resultant signal of the entire array. The signal detection method based on the MSBA can obtain a better output in high frequency signal detection under heavy noise. This method is verified and confirmed by numerical simulation and a practical case for mechanical fault diagnosis.

This chapter is organized as follows. In Sect. 2, bistable SR model is presented. In Sect. 3, normalized scale transform are introduced and validated by simulation and experiment. In Sect. 4, SR circuit module based on parameter tuning is designed and validated by simulated and experimental signal. In Sect. 5, the MSBA model is constructed and the SR effect of the model is analyzed. The signal detection approach based on MSBA is proposed and numerical simulations are carried out, which is followed by experiment on enhanced detection of rolling element bearing. Finally, the conclusions are outlined in Sect. 6.

2 Bistable Stochastic Resonance Model

The study of stochastic resonance in signal processing has received considerable attention over the last decades. In the context, stochastic resonance is commonly described as an approach to increase the SNR at the output through the increase of the special noise level at input signal. The essence of the physical mechanism underlying classical SR has been described in [1, 5, 9].

Considering the motion in a bistable double-well potential of a lightly damped particle subjected to stochastic excitation and a harmonic excitation (i.e., a signal) with low frequency ω_0. The signal is assumed to have small enough amplitude that, by itself (i.e., in the absence of the stochastic excitation), it is unable to move the particle from one well to another. We denote the characteristic rate, that is, the escape rate from a well under the combined effects of the periodic excitation and the noise, by $\alpha = 2\pi n_{\text{tot}}/T_{\text{tot}}$, where n_{tot} is the total number of exits from one well

during time T_{tot}. We consider the behavior of the system as we increase the noise while the signal amplitude and frequency are unchanged. For zero noise, $\alpha = 0$, as noted earlier. For very small noise, $\alpha < \omega_0$. As the noise increases, the ordinate of the spectral density of the output noise at the frequency ω_0, denoted by $\Phi_n(\omega_0)$, and the characteristic rate α increases. Experimental and analytical studies show that, until $\alpha \approx \omega_0$, a cooperative effect (i.e., a synchronization-like phenomenon) occurs wherein the signal output power $\Phi_s(\omega_0)$ increases as the noise intensity increases. Remarkably, the increase of $\Phi_s(\omega_0)$ with noise is faster than that of $\Phi_n(\omega_0)$. This results in an enhancement of the SNR. The synchronization-like phenomenon plays a key role in the mechanism as described in [23].

At present, the most common studied SR system is bistable system, which can be described by the following Langevin equation

$$\dot{x} = ax - bx^3 + A\sin(\omega_0 t + \varphi_0) + \Gamma(t) \tag{1}$$

where $\Gamma(t)$ is noise term and $<\Gamma(t), \Gamma(0)> = 2D\delta(t)$, $A\sin(\omega_0 t + \varphi_0)$ is a periodic driving signal. Generally, it is also written as the form of Duffing equation

$$\ddot{x} = -\beta\dot{x} + ax - bx^3 + A\ \sin(\omega_0 t + \phi_0) + \Gamma(t) \tag{2}$$

where β is the damping coefficient.

3 Normalized Scale Transform

3.1 Basic Theory of Normalized Scale Transform

Equation (1) has two stable solutions $x_s = \pm\sqrt{a/b} = \pm c$ (stable points) and a unstable solution $x_u = 0$ (unstable point) when $A = D = 0$, here potential of Eq. (1) is given by

$$V(x) = -\frac{1}{2}ax^2 + \frac{1}{4}bx^4 \tag{3}$$

The height of potential is

$$\Delta V = V(0) - V(c) = \frac{a^2}{4b} \tag{4}$$

When adding the modulation signal, potential function is

$$V(x,t) = -\frac{1}{2}ax^2 + \frac{1}{4}bx^4 - Ax\cos\omega_0 t \tag{5}$$

For a stationary potential, and for $D \ll \Delta V$, the probability that a switching event will occur in unit time, i.e. the switching rate, is given by the Kramers formula [2]

$$r_0 = (2\pi)^{-1}[V''(0)|V''(c)]^{1/2}\exp(-\Delta V/D) \tag{6}$$

where $V''(x) \equiv \mathrm{d}^2V/\mathrm{d}x^2$. We now include a periodic modulation term $A\ \sin\omega_0 t$ on the right-hand-side of (1). This leads to a modulation of the potential (5) with time: an additional term $-Ax\ \cos\omega_0 t$ is now present on the right-hand-side of (5). In this case, the Kramers rate (6) becomes time-dependent:

$$r(t) \approx r(0)\exp(-Ax\ \sin\omega_0 t/D) \tag{7}$$

which is accurate only for $A \ll \Delta V$ and $\omega_0 \ll \{V''(\pm c)\}^{1/2}$. The latter condition is referred to as the adiabatic approximation. It ensures that the probability density corresponding to the time-modulated potential is approximately stationary (the modulation is slow enough that the instantaneous probability density can 'adiabatically' relax to a succession of quasi-stationary states). The slow modulation means the signal to detect is confined to a rather low frequency range and small amplitude, and theoretical analysis and deduction are also based on this hypothesis. As we all know, the characteristic frequency reflecting mechanical system state exceeds the range of limit, so how to detect the high frequency signal is of great importance in weak characteristic signal detection of mechanical system. Here we proposed a kind of transform to solve the problem.

Considering the bistable system modeled by Eq. (1), where A is amplitude of the input signal, $\omega \gg 1$ is its frequency, $\Gamma(t)$ is Gaussian white noise with the correlation $\langle\Gamma(t)\rangle = 0$; $\langle\Gamma(t), \Gamma(0)\rangle = 2D\delta(t)$, and D is the noise intensity, when a and b are positive real numbers, take the variable substitutions

$$z = x\sqrt{b/a}, \quad \tau = at \tag{8}$$

Substituting Eq. (8) into Eq. (1), we can obtain

$$a\sqrt{\frac{a}{b}}\frac{dz}{dt} = a\sqrt{\frac{a}{b}}z - a\sqrt{\frac{a}{b}}z^3 + A\ \cos\left(\frac{\omega_0}{a}\tau + \phi_0\right) + \Gamma\left(\frac{\tau}{a}\right) \tag{9}$$

where the noise $\Gamma(\tau/a)$ satisfies $\langle\Gamma(\tau/a)\Gamma(0)\rangle = 2Da\delta(\tau)$. Therefore

$$\Gamma\left(\frac{\tau}{a}\right) = \sqrt{2Da}\xi(\tau) \tag{10}$$

where $\langle\xi(\tau)\rangle = 0$, $\langle\xi(\tau), \xi(0)\rangle = \delta(\tau)$.

Substituting Eq. (10) into Eq. (9), then

$$a\sqrt{\frac{a}{b}}\frac{dz}{dt} = a\sqrt{\frac{a}{b}}z - a\sqrt{\frac{a}{b}}z^3 + A\cos\left(\frac{\omega_0}{a}\tau + \phi_0\right) + \sqrt{2Da}\xi(\tau) \tag{11}$$

Equation (11) can be simplified into

$$\frac{dz}{dt} = z - z^3 + \sqrt{\frac{b}{a^3}}A\cos\left(\frac{\omega_0}{a}\tau + \phi_0\right) + \sqrt{\frac{2Db}{a^2}}\xi(\tau) \tag{12}$$

Equation (12) is a normalized form and equals to Eq. (1). The frequency of the signal after the transform is $1/a$ times of which before transform. Hence, through the chosen of larger parameter a, high frequency signal can be normalized to low frequency to satisfy the request of the theory of SR.

During the numerical simulation, the variance σ^2 is used to describe the statistical property of the white noise. As the noise intensity D is influenced by sample step h, the actual value $D = \sigma^2 h/2$.

Considering the RMS of the noise is $\sigma_0 = \sqrt{2D/h}$ before transform, after the transform, the intensity of the noise changed to $2Db/a^2$. And because the sample frequency descends, the sample step becomes a times of the original sample step. Therefore, the RMS of the noise after transform is $\sigma = \sqrt{2Db/(a^2 \cdot ah)}$. The ratio of the noise RMS after the transform to which before the transform is

$$\sigma/\sigma_0 = \sqrt{b/a^3} \tag{13}$$

It is easy to be seen that, after the transform, the signal and noise are amplified $\sqrt{b/a^3}$ times.

3.2 Simulation Result of Normalized Scale Transform

In the following, the scale transform will be validated through a numerical simulation. We passed the mixed signal through the model of bistable system with parameters $a = b = 1$, $A = 0.3$, $f = 0.01$ Hz, $\sigma = 1.2$, and analyze the spectrum of the output signal. Figure 1a, b shows the mixed signal and its spectrum, while Fig. 1c, d gives the output of the bistable system and the spectrum of the output signal. From Fig. 1d, it can be seen that although the input SNR $= 20\log(A/\sigma) = -12.04$ dB, there is a clear spectrum line at $f = 0.01$ Hz, and the noise fades obviously.

If the signal frequency is changed to $f = 1$ kHz, according to the transform principle, we can take the parameters $a = b = 10^5$. Conditioned the mixed signal through the SR model, the result can be shown in Fig. 2. The detection result based on the normalized scale transform is shown in Fig. 2. Figure 2a, c are the

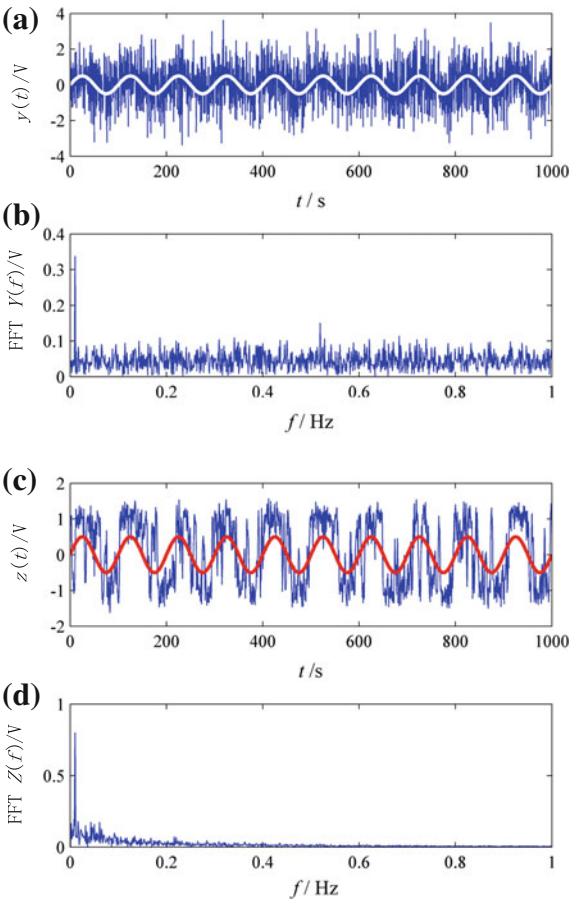


Fig. 1 Time-domain and its FFT of the input and output when $f = 0.01$ Hz. **a** and **b**: the input; **c** and **d**: the output by one-time

waveforms of input and output signals. Figure 2b, d are their FFT spectra. The component of 1 kHz is revealed clearly in Fig. 2d.

The signal and spectrum in Fig. 2 is consistent with Fig. 1, and the only difference is time domains and frequency coordinates of the spectrum. The noise components are greatly suppressed, and the detecting signal is standing out, which shows that the transform method is suitable to the detection of high frequency signal.

3.3 Application of Normalized Scale Transform

In cases where it is desired to process sampled discrete vibration signals, we realized that it would be possible to enhance the bearing characteristic components using SR method. As mentioned above, the SR normalized scale transform is

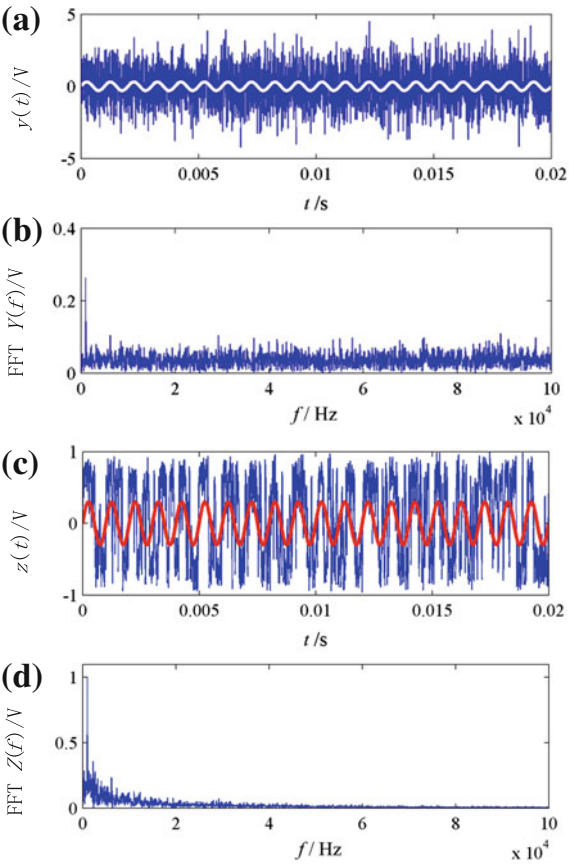


Fig. 2 Time-domain and its FFT of the input and output when $f = 1$ kHz. **a** and **b**: the input; **c** and **d**: the output by one-time

suitable for big parameter signal processing. In this section, SR normalized scale transform is applied in bearing fault diagnosis. The schematic diagram of bearing fault enhanced detection method is shown in Fig. 3. After sampling, the analog vibration signal is converted to input data as depicted in Fig. 3. Then, band-passed vibration signal is demodulated based on Hilbert transformation, and the output is bearing vibration envelope signal. The band-pass filter parameters are set to cover bearing natural resonance frequencies. Finally, envelope signal enhanced by SR normalized scale transform is transformed to frequency domain through FFT algorism and fault features are extracted. The procedures of this method from input vibration data to fault features are carried out by software, in other words, achieved by computation.

This method based on SR normalized scale transform is applied to vibration signal from machinery fault simulation test rig shown in Fig. 4. Tests were carried out on the test rig with normal and planted-in inner fault bearings. The rig is driven by a variable-speed electric motor. For these tests, the shaft speed is 628 r/min with two rotor disks on the shaft. The Bearing1 in Fig. 6 is alternated with normal

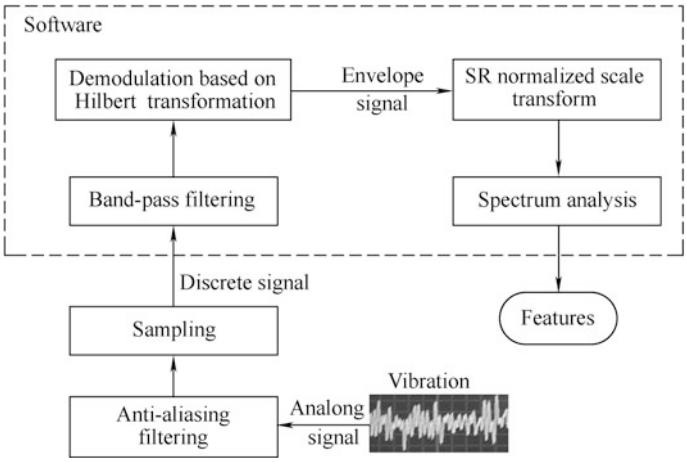


Fig. 3 Schematic diagram of enhanced bearing fault detection method using SR normalized scale transform

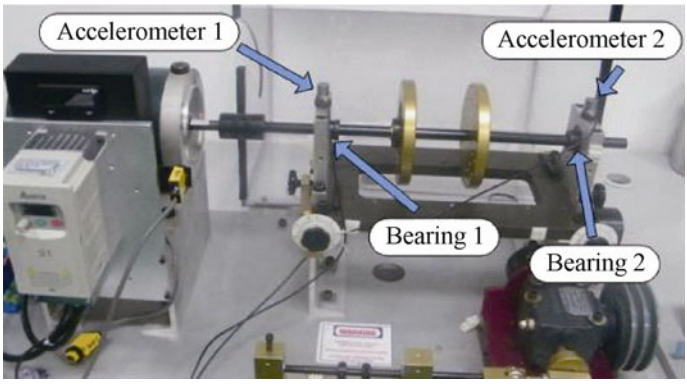


Fig. 4 Machinery fault simulation test rig

bearing, bearing with 0.2 mm planted inner race fault and bearing with 0.5 mm planted inner race fault, which are shown in Fig. 5. Signals were measured by an accelerometer on the casing immediately above it. Details of the geometry of the bearings are shown in Table 1. The expected inner race fault frequency (Ball pass frequency, inner race, BPFI) is 70.28 Hz. The raw vibration data was collected with the sampling rate 50 kHz. And the collected data length is 1 s. Figure 6 displays the recorded raw time signals from Accelerometer1 denoted in Fig. 6: (1) normal bearing, (2) bearing with 0.2 mm inner race fault and (3) bearing with 0.5 mm inner race fault.

From the raw signal we can see that there are more impacts in the vibration signals of 0.2 and 0.5 mm inner race fault than the signal of normal bearing.

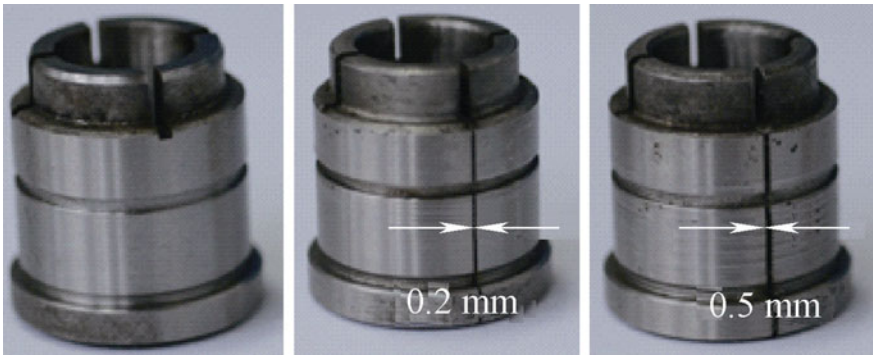


Fig. 5 Inner races of normal, with 0.2 mm planted fault and with 0.5 mm planted fault bearing (from *left* to *right*)

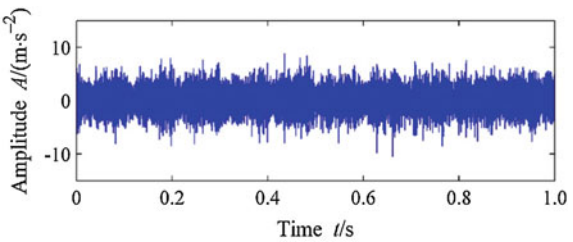
Table 1 Test bearing parameters

Parameter	Value
Roller diameter d/mm	7.50
Pitch diameter D/mm	34.50
Roller number n	11
Contact angle ϕ/(°)	0
Shaft speed v/(r · min^{-1})	628

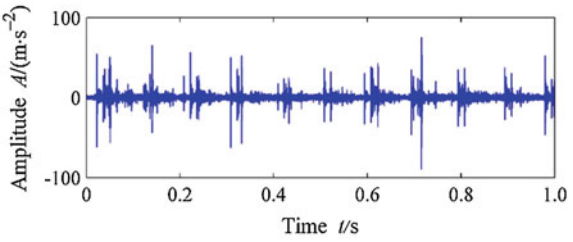
And the signal of 0.2 mm inner race fault contains obvious periodic impacts at about shaft speed, as shown in Fig. 6b. This should be caused by imbalance of the rotor system. However, we could not make sure whether there is local damage on any of the bearing component or not.

The signals were demodulated on frequency range from 8 to 12 kHz, which covers one of the bearing nature resonance bands. And Fig. 7 shows the envelope spectra of the three cases, which is up to 500 Hz—the band dominated by shaft speed, bearing fault characteristic component and their harmonics. The BPFI and its harmonics are indicated by harmonic cursors in the envelope spectra. It can be seen in Fig. 7a that there are only shaft speed component and its second harmonic. In the envelope spectrum of 0.2 mm inner race fault shown in Fig. 7b, discrete spectrum components including shaft speed, BPFI and their harmonics can be seen, but the BPFI and the second harmonic is not clear. However, the BPFI and its harmonics are obvious in Fig. 7c, since 0.5 mm inner race fault is rather severe. We use local signal to noise ratio (LSNR) as the indicator of the BPFI component, which is defined as

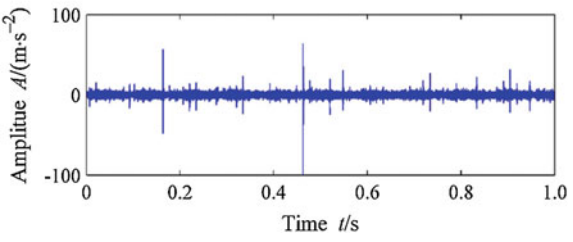
$$R = 10\lg\left\{\lim_{\Delta f\to 0}\left[\int_{f-\Delta f}^{f+\Delta f}(S(f)/S_N(f))\mathrm{d}f\right]\right\} \tag{14}$$



(a) Normal bearing



(b) Bearing with 0.2mm inner race fault



(c) Bearing with 0.5mm inner race fault

Fig. 6 Raw vibration signals of the experiment

where $S(f)$ denotes the power density at signal frequency f, $S_N(f)$ is the noise mean power density around f. The LSNR indicators of envelope spectra of the three cases are 8.88, 8.58 and 14.01 dB respectively. We could not distinguish the 0.2 mm inner race fault bearing from normal bearing only by the envelope spectrum.

Figure 8 shows the corresponding spectra of the three signals after the vibration data are processed using the method shown in Fig. 3. The SR system parameters were tuned according to a target signal frequency of 200 Hz. It is found that the inner race fault component of 0.2 mm inner race fault was enhanced greatly, but the corresponding components of the normal bearing did not show up. The LSNR indicators of normal bearing and 0.2 mm inner fault are 8.58 and 12.22 dB. However, we could not see obvious change at the inner race fault component of the 0.5 mm inner fault case, and the LSNR indicator increases slightly to 14.16 dB. The shaft speed and its second harmonic were enhanced simultaneously in the three cases.

Although effective in the application of sampled signals processing, due to the fact that it is realized by software calculation, the normalized scale transform has

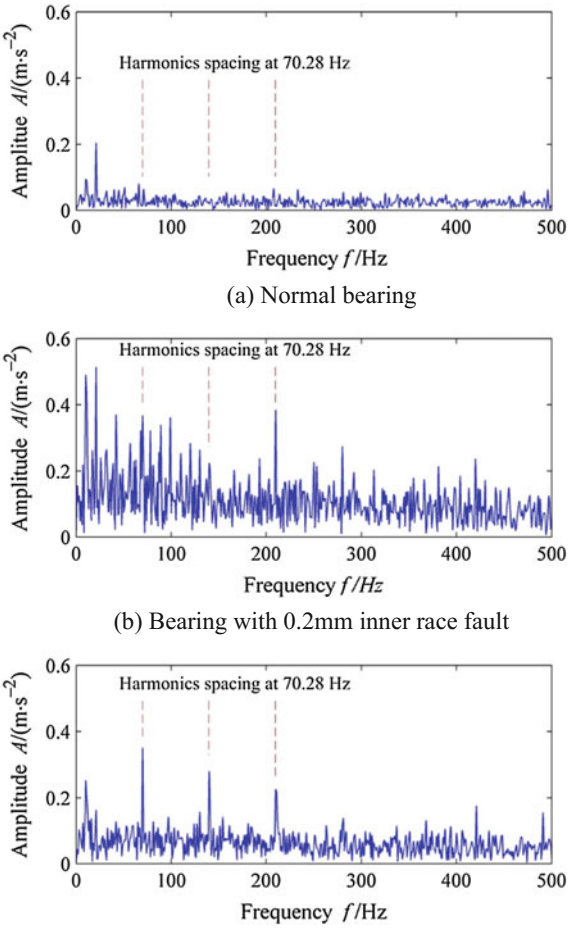


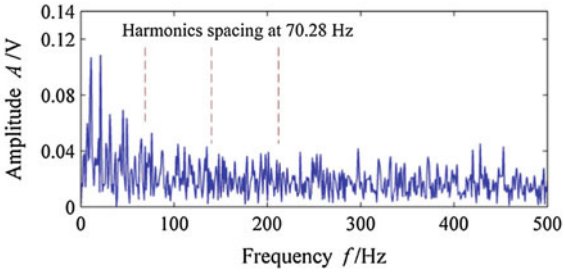
Fig. 7 Envelope spectra of the test bearings

some drawbacks: (1) The sampling frequency should be much higher than the Nyquist frequency to make sure that the target signal is in the low area of whole frequency range; (2) A lot of computation is needed to obtain the solution of the differential equation.

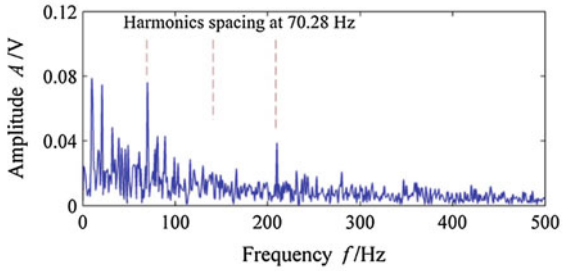
4 SR Circuit Module

4.1 Circuit Module

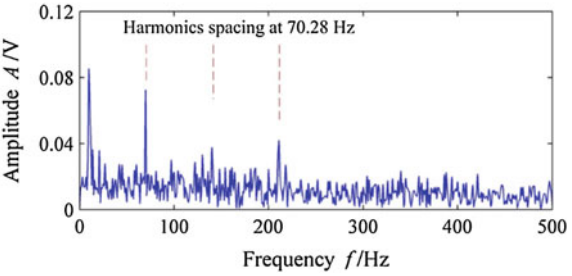
Because software realization of SR requires intensive computation and high sampling frequency, it would be a practical way to actualize SR by using hardware



(a) Normal bearing



(b) Bearing with 0.2mm inner race fault



(c) Bearing with 0.5mm inner race fault

Fig. 8 Enhanced spectra of SR normalized scale transform

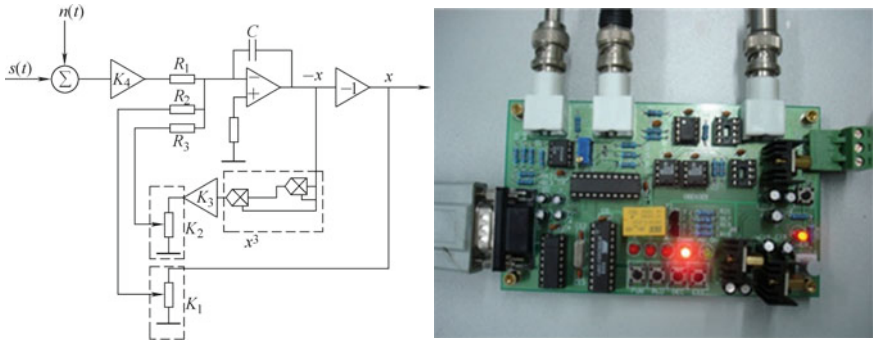
devices. To save the computational resource and the SR processing time, a circuit module is designed in this section.

The integral form of Eq. (1) is

$$x(t) = \int \left[ax - bx^3 + s(t) + \Gamma(t)\right]\mathrm{d}t \tag{15}$$

where $s(t)$ is the signal to be detected. Equation (15) could be expressed as a nonlinear system with a feedback loop, which involves amplifier, integrator, multiplicator etc. The feedback loop could be realized by amplifier, resistance and capacitances. Figure 9 is the frame and concrete SR circuit module.

According to the circuit principles, the mathematical model (nonlinear stochastic integral equation) of the circuit module can be written as



(a) Circuit diagram frame (b) Circuit photo

Fig. 9 Design of the SR circuit module

$$x = \int \left\{ \frac{K_4[s(t) + \Gamma(t)]}{R_1 C} - \frac{K_2 K_3 x^3}{R_3 C} + \frac{K_1 x}{R_2 C} \right\} \mathrm{d}t \tag{16}$$

The differential form of Eq. (15) is

$$\dot{x} = \frac{K_4[s(t) + \Gamma(t)]}{R_1 C} - \frac{K_2 K_3 x^3}{R_3 C} + \frac{K_1 x}{R_2 C} \tag{17}$$

By comparing Eq. (17) and Eq. (1), the circuit system parameters can be written as $a = K_1/R_2C$, $b = K_2K_3/R_3C$, $A = K_4/R_1C$, $\Gamma'(t) = A\Gamma(t)$. The two stable status of the bistable circuit model, i.e., the two penitential wells' locations are

$$x_{1,2} = \pm\sqrt{\frac{a}{b}} = \pm\sqrt{\frac{K_1 R_3}{K_2 K_3 R_2}} \tag{18}$$

The system potential is

$$V(x) = -\frac{a}{2}x^2 + \frac{b}{2}x^4 = -\frac{1}{2}\left(\frac{K_1}{R_2 C}\right)x^2 + \frac{1}{4}\left(\frac{K_2 K_3}{R_3 C}\right)x^4 \tag{19}$$

So, the potential height is

$$\Delta V = \frac{a^4}{4b} = \frac{1}{4}\frac{K_1^2 R_3}{R_2^2 C K_2 K_3} \tag{20}$$

Equation (17) is consistent with Eq. (1) formally and intrinsically. According to bistable stochastic resonance system theory, parameter a correlates with signal frequency, and b influences ΔV. The circuit module is physically coincident with the bistable model in theory.

The parameters of the circuit R_1, R_2 and R_3 are 10 kΩ, C = 150 pF, K_3 = 0.01. Given the other parameters, adjusting K_1 could tune a (0–666,667) to adapt to signal frequency, and adjusting K_2 could tune b (0–6667) to adapt to different noise intensity. Via adjusting resistance coefficients K_1, K_2 or both, potential height and stable status could also be tuned. Parameters tuning can be realized by adjusting the two resistances of the circuit module designed in this section. The difference is that the input signal amplitude should be retuned based on parameters of the SR module. The weak target signal will be revealed, when signal, noise and nonlinear system are matched. Since both the input and output of circuit module are both analog signals, sampling frequency of the output signal is just demanded to catch the signal to be revealed. In other words, the high sampling frequency could be avoided. This will be validated by the simulation test in the next section.

4.2 Simulated Experiment of Circuit Module

The input signal is a sinusoidal signal mixed with a white noise generated by generators. The signal frequency is 10 kHz and the amplitude is 0.04 mV. The white noise intensity RMS is 0.6 mV. Then the input signal SNR is −23.52 dB. The input signal waveform and its spectrum are shown in Fig. 10a, b.

Adjusting circuit module coefficients K_1 to 2/63 and K_2 to 50/63, that is to say, the system parameters are a = 21,164, b = 5291 and the stable states are $x_{1,2} = \pm 2$ V. The highest frequency, which could be enhanced theoretically, is calculated to 2116.4 Hz. The output signal waveform and its spectrum are shown in Fig. 10c, d. The 10 kHz signal is revealed clear in the spectrum. The signal sample frequency is 100 kHz, and the data length is 2000.

4.3 Application of Circuit Module

If SR is realized by circuit module, it would be possible to replace the SR normalized scale transform with circuit module and then change the input data to analog signal. As mentioned in Sect. 3, the sampling frequency, which could catch the signal interested under Nyquist sampling law, would be adequate for SR circuit module output signal. Moreover, there is no need to sample the signal at the beginning, if the signal processing procedures before SR circuit module are implemented by hardware. The bearing fault enhanced detection method using SR circuit module is shown schematically in Fig. 11. The parameters of SR circuit module are tuned according to the signal interested. The analog vibration signal from bearing is filtered by a band-pass filter directly. Then, the band-passed vibration signal is demodulated by envelope detection. The band-pass filter parameters are set to cover the bearing nature resonance frequency band. Envelope signal is enhanced by SR circuit module and then transformed to frequency domain

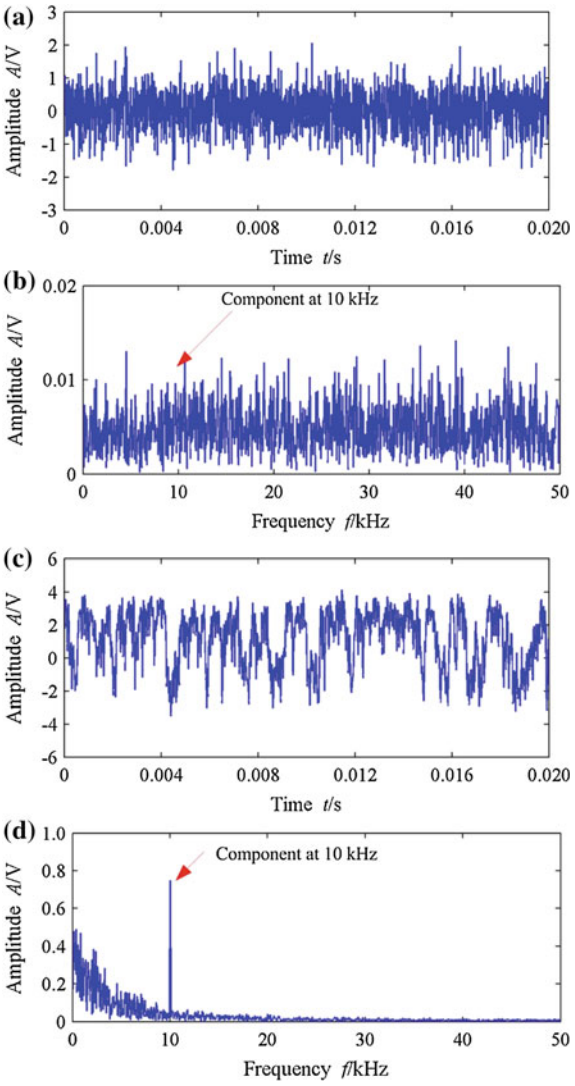


Fig. 10 Detection of 10 kHz signal by SR circuit module

through FFT algorithm after signal sampling at a much lower rate than software method. The signal processing procedures before sampling are all realized by hardware.

This method based on SR circuit module is applied to vibration signal shown in Fig. 6. Then, the analog vibration signals of the three cases were processed by hardware with SR circuit module according to the diagram shown in Fig. 11. Adjusting circuit module coefficients K_1 to 2/63 and K_2 to 50/63, which means that the system parameters are $a = 21{,}164$, $b = 5291$. The output signal of one second was collected at sampling rate 1 kHz. The FFT spectra of the three cases are shown

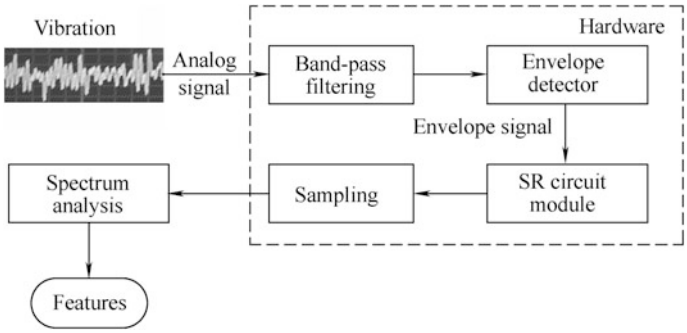


Fig. 11 Schematic diagram of enhanced bearing fault detection method using SR circuit module

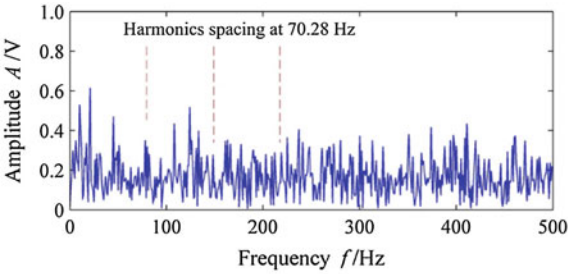
in Fig. 12. Similar results to Fig. 10 were obtained with the LSNR indicators of 8.11, 12.23 and 13.67 dB. The shaft speed and its second harmonic were also enhanced simultaneously in the three cases.

5 Multi-scale Bistable Array SR

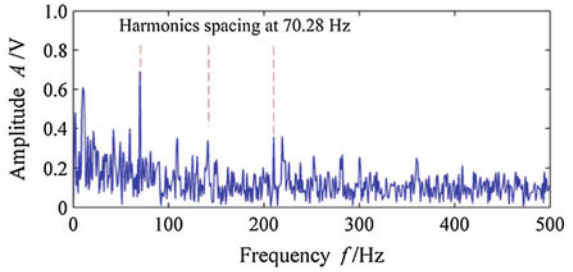
5.1 Stochastic Resonance Effect in MSBA

The SR effect is still shown in a nonlinear bistable system when the white noise is changed to band-limited noise, which indicates that it is possible to realize the SR by tuning the band-limited noise [5, 24]. To improve the signal processing based on SR when the original noise intensity is beyond the optimal level, the input signal is decomposed into multi-scale signals by orthogonal wavelet transform. Stationary white noise with zero mean can be decomposed into independent band-limited noises by orthogonal wavelet transform. The reconstructed detail at each scale and the approximation signal at the last scale are independent of each other owing to the orthogonality of wavelet base.

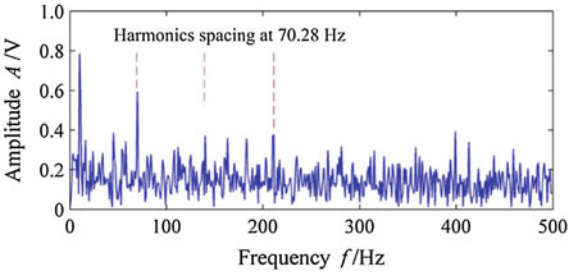
The SR effect of the bistable model of Eq. (1) is investigated using a sine signal plus a single scale noise as the input signal. By adjusting the noise intensity at each scale, Fig. 13 illustrates the SR enhancement effect of each scale noise, which indicates that SR can also be produced by each scale noise alone. The signal amplitude $A_0 = 0.3$, frequency $f_0 = 0.01$ Hz, system parameters $a = b = 1$, sampling frequency $f_s = 5$ Hz, and data length $N = 4000$. In the context, a_j and d_j denote reconstructed approximation signals and detail signals, respectively for convenience. It can be seen that the scale noise a_3 has the effect similar to the white noise, and the other scale noises still show clear SR effect when taking higher noise intensity. In a bistable system, the output SNR curves produced by different scale noises show dissimilar SR mechanisms. Now, an interesting question arises,



(a) Normal bearing



(b) Bearing with 0.2mm inner race fault



(c) Bearing with 0.5mm inner race fault

Fig. 12 Enhanced spectra of SR circuit module

namely, can we further improve the output SNR under large noise intensity? The answer is positive and lies in the different SR effect of each scale noise.

By combining uncoupled parallel array of dynamical subsystems with colored noise SR effect, an MSBA consisting of bistable units formulated as Eq. (1) is constructed. Figure 14 illustrates the configuration of the MSBA. The input signal is decomposed into different scale signals by wavelet transform. The driving signal of the MSBA, which is in the low frequency region, is supposed to be contained in the approximate signal a_J. The approximate signal a_J and each scale noise d_j reconstruct the new input signal of each bistable element. Then, the number of array elements is equal to the scale number J. Being uncoupled between any two elements, the outputs of all units are averaged together to produce the entire array output $y(t)$. Similar to uncoupled parallel SR array, each element is subjected to an independent array noise d_j and the same noisy input signal a_J. However, the inner

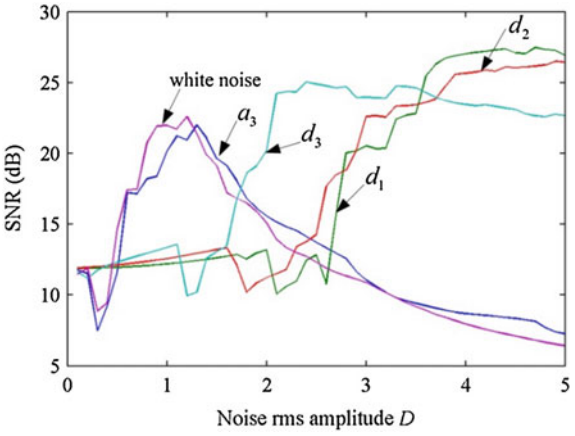


Fig. 13 Output SNR curves of SR produced by individual scale noises alone

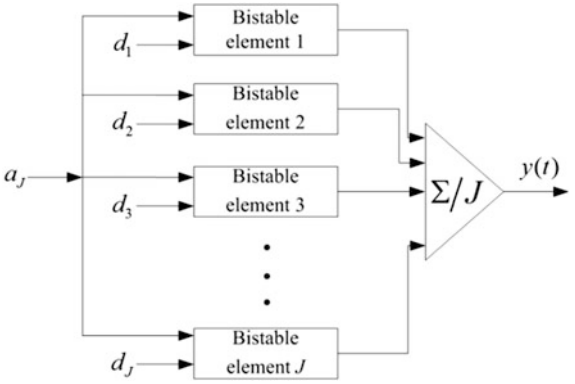


Fig. 14 MSBA model of J elements

array noise intensity and characteristics of the MSBA are different from those of the uncoupled parallel array. The noise intensity of input signal is reduced after decomposition. The noise at different scale has a different contribution to the SR effect of the single bistable element in the array.

Numerical simulation is performed according to the MSBA system given in Fig. 14. Each bistable element in the MSBA is formulated as Eq. (1). The SR effect of the MSBA is evaluated with tuning the input noise intensity D and the analyzed scale number J. The other system parameters are chosen as $a = b = 1$, $A_0 = 0.3$, $f_0 = 0.01$ Hz and $\varphi = 0$. The sampling frequency is set to be $f_s = 500f_0 = 5$ Hz, and the data length of the input signal is set to be 4000.

Figure 15 shows the SR effect of the MSBA with tuning noise intensity D and the analyzed scale number J. The array output SNR curves, from bottom up, correspond to $J = 1, 2, 3, 4, 5$, and 6, respectively. The other parameters are chosen as $a = b = 1, A_0 = 0.3, f_0 = 0.01$ Hz and $\varphi = 0$. The results indicate that the tuning noise intensity D of the input signal produces an obvious SR effect on the MSBA.

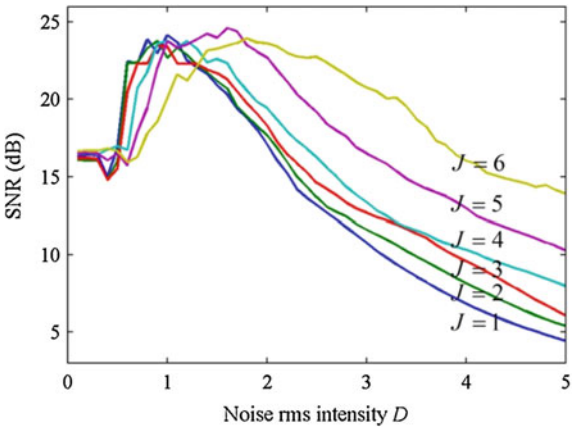


Fig. 15 Output SNR of the MSBA versus tuning noise intensity D and analyzed scale number J

When the value of D increases, the array output SNR first increases and then decreases after reaching a maximum. It can be found that the output SNR resonance region is broadened with the increase of the analyzed scale number J, and also the maximum point is moved to a larger noise intensity D with almost the same value. When the noise intensity D becomes larger, the SNR curves of $J > 1$ will go up compared with the curve of $J = 1$. This means that the output signal of the MSBA has been enhanced further than that of the single bistable system at a larger input noise intensity D. Thus, the model of MSBA has admirable capability in signal processing based on SR under large noise.

The noise intensity curves of the MSBA output signal versus input noise intensity D and analyzed scale number J are shown in Fig. 16. The noise intensity curves, from top down, correspond to $J = 1$, 2, 3, 4, 5, and 6, respectively. The other simulation parameters are the same as in Fig. 15. The output signal is filtered by a high pass filter, which is cut off at 0.1 Hz, to eliminate the driving frequency component. It can be seen that the output noise intensity is reduced gradually when the analyzed scale number J increases for a given input noise intensity. This indicates that the MSBA can achieve a better signal quality than that obtained by the conventional single SR unit.

Figure 17 compares the signal detection result of the conventional single SR unit with that of the MSBA at fixed noise intensity. The analyzed scale is $J = 6$ and the other parameters are the same as in Fig. 15. The target signal is submerged in the heavy noise ($D = 2.3$, $A_0 = 0.3$) as seen in the input signal wave in Fig. 17a. The dashed line and right y-axis in (a), (b) and (c) show the input target signal.

Comparing the output of the conventional single SR unit in Fig. 17b with that of MSBA in Fig. 17c, we can find that the MSBA can obtain smoother output waveform and lower noise. Sub-figures (d), (e) and (f) are the spectrums of sub-figures (a), (b) and (c), respectively. The above study shows that the proposed MSBA model has the capability of detecting signal under heavy noise background and can obtain the output signal with lower noise intensity correspondingly.

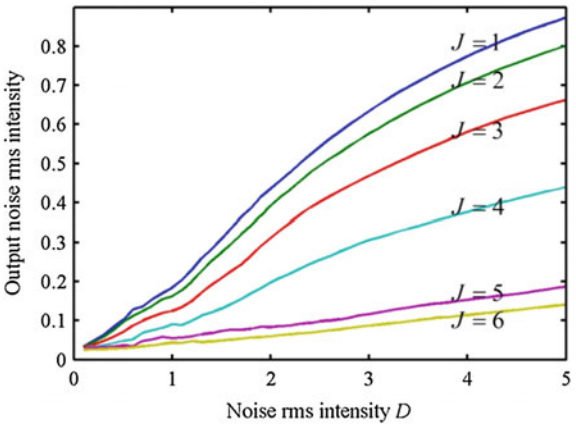


Fig. 16 Noise intensity curves of the MSBA output signal versus input noise intensity and analyzed scale number

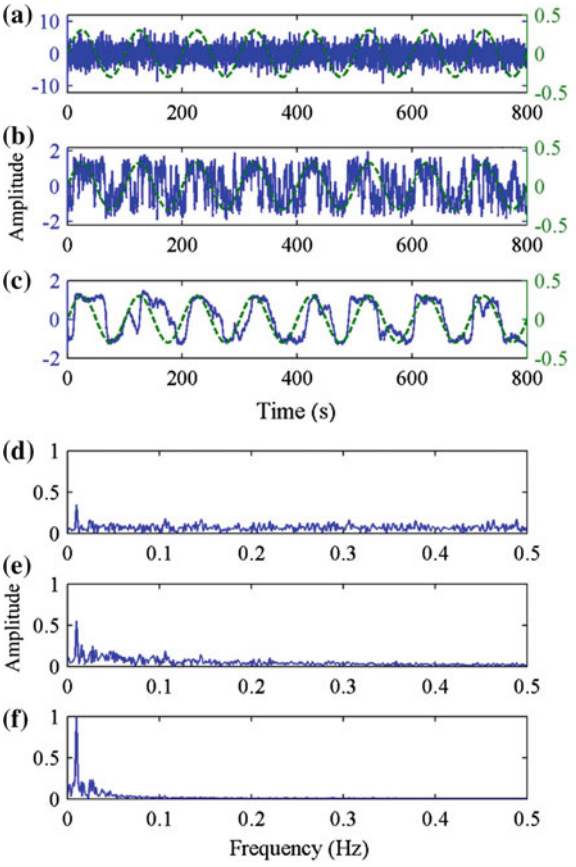


Fig. 17 Comparison of signal detection between conventional model and MSBA model

5.2 Signal Detection and Numerical Simulation

Based on the MSBA model and normalized scale transform, we present a novel weak signal detection method. The framework of the method is shown in Fig. 18. The first part is signal decomposition based on wavelet transform. a_J often contains low frequency interference which will be also enhanced by SR in practical application. The selection principle of the decomposition level J is that the d_J should cover the frequency of the target signal. Then, scale d_J is processed as the input signal of the MSBA model. The other scales of high frequency noises, as inner noises of the array, are inputted to the bistable units of MSBA, respectively. The second part is tuning MSBA parameters (a, b, k) using normalized scale transform for high frequency target signal detection, where k is the amplitude coefficient of the input signal. Finally, output signals of the units after parameter tuning are summed up and divided by the array size to obtain the resultant signal $y(t)$.

High frequency signal detection based on SR can be carried out by normalized scale transform. The SR effect of MSBA has been validated by simulation in Sect. 5.1. However, the whole enhancing effect of the combination of normalized scale transform and MSBA on weak signal still demands further verification.

Normalized scale transform consists of two parts. One is system parameter tuning for high frequency signal processing, the other is input signal amplitude tuning for output optimization. Firstly, the effect of normalized scale transform on MSBA for high frequency signal detection is illustrated by simulation. The MSBA output SNR curves of different frequency signals are depicted in Fig. 19. The tuft of six SNR curves (solid line) is corresponding to target signal frequencies f_0 = 0.01, 0.1, 1, 10, 100 and 1000 Hz, respectively. The sampling frequency is set to be f_s = 500 f_0. $a = b = f_0/0.01$, other parameters are set as the same as in Fig. 15. The effect of normalized scale transform on a single bistable unit is also shown in Fig. 19 for comparison. The tuft of six SNR curves (dashed line) is corresponding to the same target signal frequency as the MSBA model, where the parameters of the single unit are the same as the parameters of MSBA. By normalized scale transform, the target signal of different frequencies can be enhanced by SR effect in the MSBA model and the single bistable model.

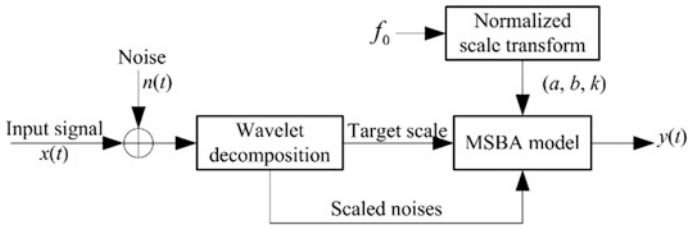


Fig. 18 Scheme of weak signal detection based on MSBA and normalized scale transform

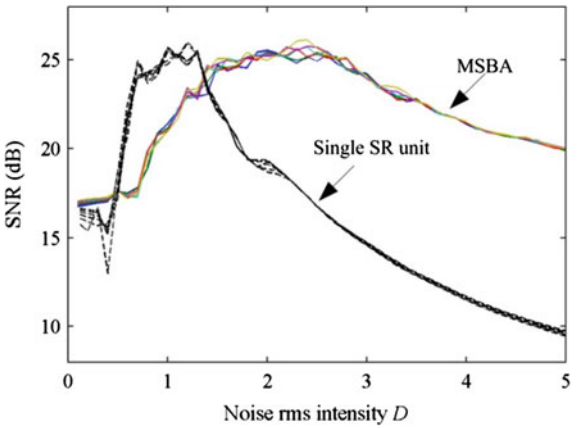


Fig. 19 Effect of normalized scale transform on high frequency signal detection

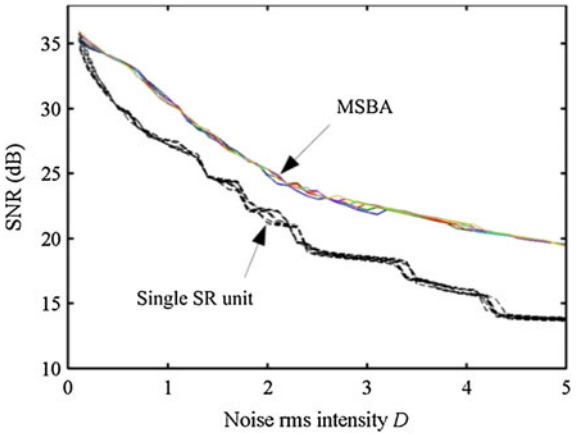


Fig. 20 SNR curves corresponding to different frequencies after amplitude tuning

Then, the effects of input signal amplitude tuning on conventional SR and MSBA are illustrated by simulation. The output SNR curves of conventional SR model (dashed line) and MSBA (solid line) after amplitude tuning are plotted in Fig. 20. The signal and system parameters used in the simulation are set to be the same as those in Fig. 19. The six curves in each tuft correspond to the target signal frequencies f_0 = 0.01, 0.1, 1, 10, 100 and 1000 Hz, respectively. Compared with the SNR curves in Fig. 19, the SNR curves after amplitude tuning are above the curves before amplitude tuning. Especially, the SNR is promoted greatly in the low noise intensity region.

A simulated 1 kHz target signal submerged in heavy noise (D = 2, A_0 = 0.3) was processed by the proposed method and the conventional SR method. The results are shown in Fig. 21. The analyzed scale is J = 6 and the other parameters are the same as in Fig. 15. The input signal wave is shown in Fig. 21a. The dashed

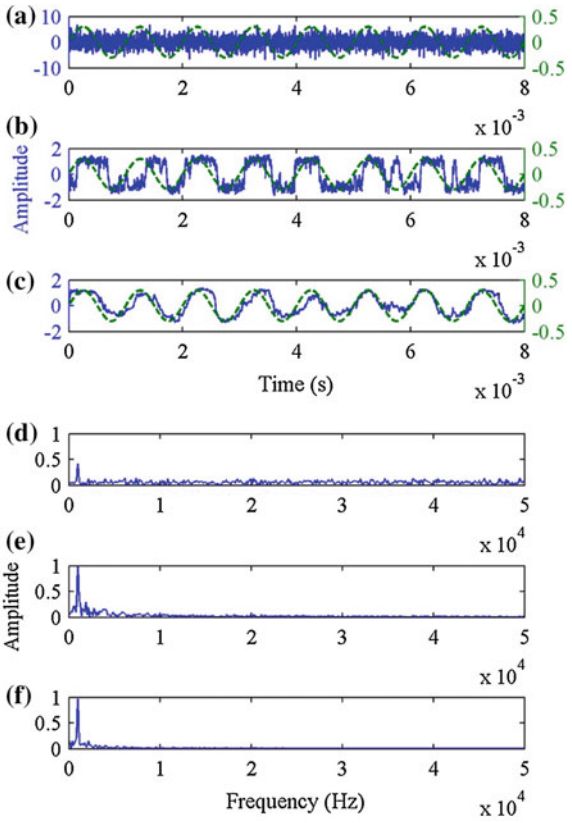


Fig. 21 Comparison of signal detection between conventional SR and the proposed method

lines and right *y*-axis in Figs. 21a–c show the input target signal. Comparing the output of the conventional single SR unit in Fig. 21b with that of the proposed method in Fig. 21c, we can find that the proposed method can detect the target signal explicitly and obtains an output signal with lower background noise. Figures 21d–f are the spectrums of Figs. 21a–c, respectively.

5.3 Application for Machinery Fault Diagnosis

A vibrational feature detection experiment for rolling element bearing incipient fault was conducted to verify the effectiveness of the proposed method. Bearings are widely used in mechanical transmission systems. Their local faults or damages usually produce characteristic frequency components, whose frequencies depend on bearing geometry, rotational speed and position of the fault.

Fault or defect is identified when the frequency component corresponding to the bearing defect induced impulses is found in the frequency domain. Then, the critical work is the weak characteristic signal detection after the demodulation of vibration impulses. In fact, vibration signatures in the envelope are taken to be the weak target signal and the noise and the other components are tuned to play an active role in the MSBA model. For the experiments in this section, the vibration signals are decomposed to make the scale d_J contain the characteristic frequency, where $J = 10$. In the following cases, the proposed method is verified in comparison with two other methods. One is the kurtogram for the detection of transient signal based on kurtosis maximization. The other is the conventional SR based on normalized scale transform.

The proposed method and the two other methods were applied to vibration signal from a test rig of machinery fault simulation, as shown in Fig. 22. Tests were carried out on the test rig with seeded inner and outer race fault bearings. The rig was driven by an electric motor with rotating speed 665 r/min. The vibration signals were collected with the sampling frequency of 50 kHz from the bearings with 0.2 mm seeded inner and outer race faults. The bearing with seeded defect was installed in the position of Bearing I during the test. The collected data time length was 1 s. The ball pass frequencies over inner and outer race defect, f_{BPFI} and f_{BPFO}, were computed to be 74.40 and 47.51 Hz.

Figure 23a displays the raw time signal collected from Accelerometer I on the test rig of the bearing with 0.2 mm inner race defect, which is denoted in the top right of Fig. 22. Figure 23b is the envelope of the signal in Fig. 23a processed by signal pre-whitening and signal demodulation [24]. The data in Figs. 23c, d are the output signals of the conventional SR method and the proposed method.

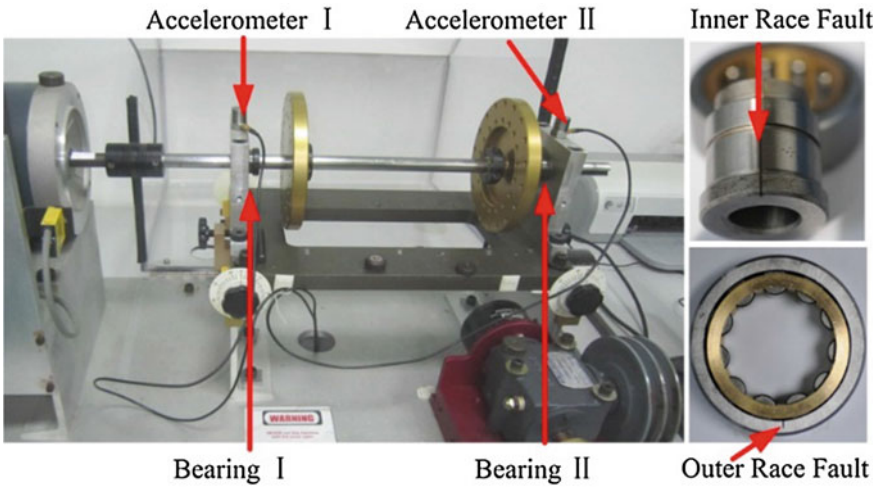


Fig. 22 Test rig and the 0.2 mm inner and outer race defect bearings

Their spectra are shown in Figs. 23e, f, respectively. It can be seen that there are some noisy impulses in the time domain signal and the envelope. The envelope signal is enhanced by traditional SR using normalized scale transform, which improves the defect identification as shown in Fig. 23e. The shaft frequency (marked as 1X) is also enhanced by the conventional SR method. The result produced by the proposed method makes the inner race fault diagnosis beyond all doubt. As seen in Fig. 23f, the characteristic component of inner race defect $f_{\mathrm{BPFI}} = 74.40$ Hz is highlighted clearly.

Figure 24 is the analyzed result of the kurtogram, where $K_{\max}$, B_{w} and f_{c} denote the maximum kurtosis, bandwidth and central frequency of the selected band, respectively. Figure 24a is the kurtogram of signal in Fig. 23a. Figure 24b is the envelope signal of the selected band, which maximizes the kurtogram. Figure 24c is the spectrum of the envelope signal. The characteristic component f_{BPFI} induced by inner race defect can be identified in the envelope spectrum. However, there are also some frequency components disturbing the inner race fault identification.

The vibration signal of 0.2 mm outer race fault is analyzed in Figs. 25 and 26 to confirm the reliability of the proposed method. Figure 25 displays results similar to that in Figs. 23 and 26 displays results similar to that in Fig. 24. All the signal and experimental parameters are set equal to those of the inner race fault identification. Only the fault type and the target signal frequency are different from the inner race

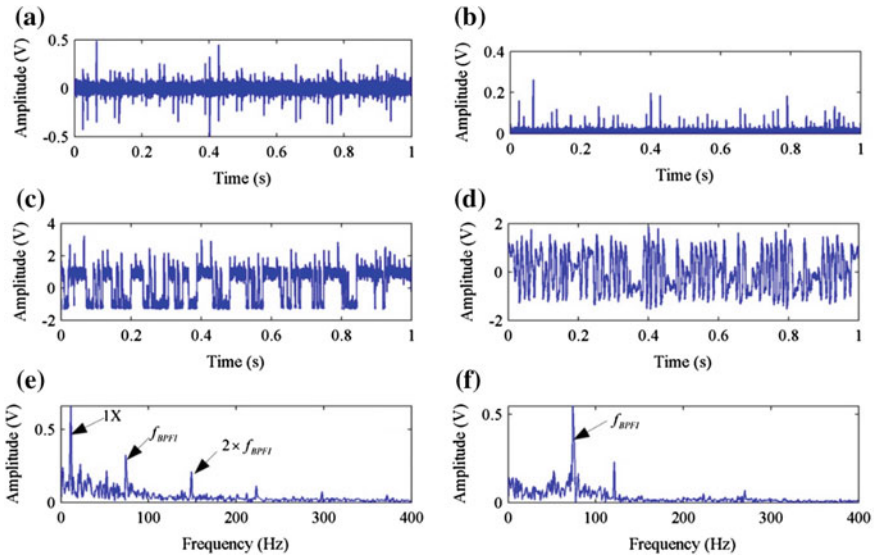


Fig. 23 Analyzed results of bearing inner race fault using the traditional SR and the proposed method

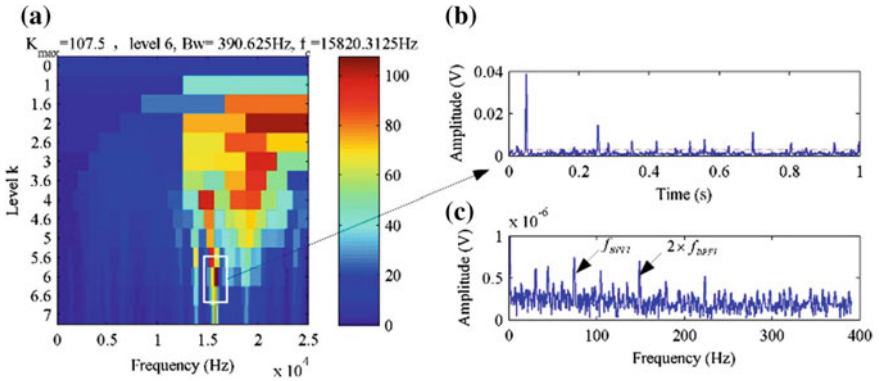


Fig. 24 Analyzed results of bearing inner race fault using kurtogram

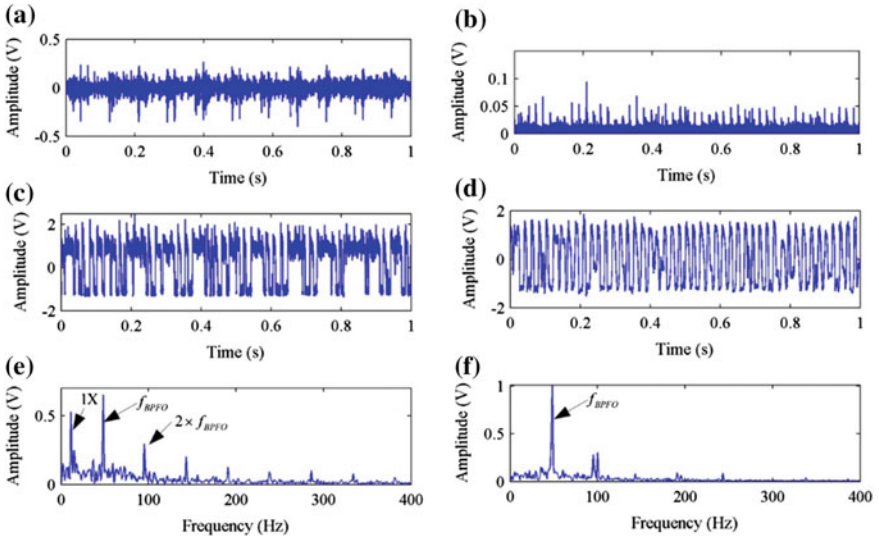


Fig. 25 Analyzed results of bearing outer race fault by using the traditional SR and the proposed method

fault identification. Similar to the result in Figs. 23 and 24, the characteristic component of outer race defect $f_{\mathrm{BPFO}} = 47.51$ Hz is highlighted clearly by the proposed method. This shows that better performance can be achieved by the proposed method in comparison with kurtogram and traditional SR method for fault diagnosis.

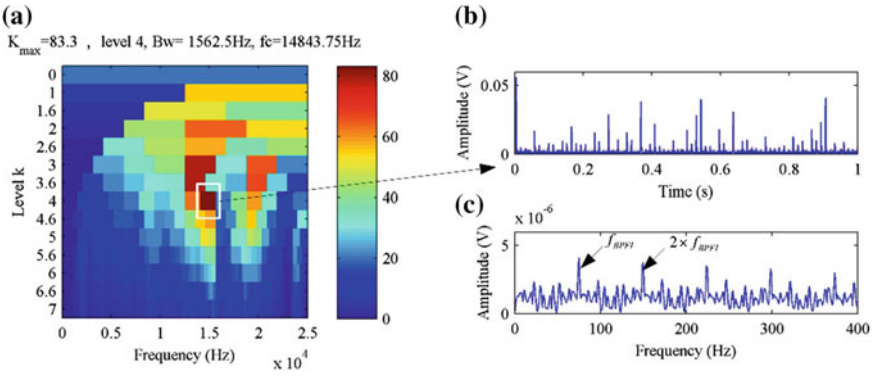


Fig. 26 Analyzed results of bearing outer race fault using kurtogram

6 Conclusions

In the chapter, we studied normalized scale transform, circuit module and multi-scale bistable array to extract characteristic fault signal of the dynamic mechanical components based on SR theory. The SR normalized scale transform is flexible and feasible for discrete signal processing, but it demands high sampling rate and expensive computation. The SR hardware module, which is suitable for processing analog signal directly, changes nonlinear system parameter tuning into resistances adjusting. The advantages of implementation by hardware are less computational task, instantaneous output, and much lower sampling frequency. The SR effect of the MSBA model could be used to detect weak signal buried by strong noise. Numerical simulation results show that the SR effect of MSBA can appear at high input noise intensity. Simulation and experiment results of the experiment on bearings with planted inner race fault demonstrate that the methods of this chapter are suitable for application in mechanical fault signal detection.

Acknowledgements The authors would like to acknowledge the support of National Natural Science Foundation of China (Grant Nos. 51475463 and 51605483) and Research Project of National University of Defense Technology (Grant No. ZK-03-14).

References

1. Benzi R., Sutera A., Vulpiana A., "The mechanism of stochastic resonance", Journal of Physics A: Mathematical and General, 1981, 14(11):L453–L457
2. Mcnamara B., Wiesenfeid K., "Theory of stochastic resonance", Physical Review A, 1989, 39(9):4854–4869
3. Jung P., Hanggi P., "Amplification of small signal via stochastic resonance", Physical Review A, 1991, 44(12):8032–8042
4. Bulsara A.R., Gammaitoni L., "Turning into noise", Physics Today, 1996, 49(3):39–45

5. Gammaitoni, L., Hanggi, P., Jung, P., et al, "Stochastic resonance", Reviews of Modern Physics, 1998, 70(1), 223–287
6. Xu B.H., Duan F.B., Bao R.H., et al, "Stochastic resonance with tuning system parameters: the application of bistable systems in signal processing", Chaos, Solitons Fractals, 2002, 13:633–644
7. Xu B.H., Li J.L., Zheng J.Y., "How to tune the system parameters to realize stochastic resonance", Journal of Physics A: Mathematical and General, 2003, 36(48):11969–11980
8. Xu B.H., Zeng L.Z., Li J.L., "Application of stochastic resonance in target detection in shallow-water reverberation", Journal of Sound and Vibration, 2007, 303:255–263
9. Hu N.Q., Chen M., Wen X.S., "The application of stochastic resonance theory for early detecting rub-impact fault of rotor system", Mechanical Systems and Signal Processing, 2003, 17(4):883–895
10. Yang D.X., Hu N.Q., "Detection of weak aperiodic signal based on stochastic resonance", In: 3rd International Symposium on Instrument Science and Technology, Xi'an: International Symposium on Instrument Science and Technology, 2004, 0210–0213
11. Zhang X. F., Hu N.Q., Cheng Z., et al, "Enhanced detection of rolling element bearing fault based on stochastic resonance", Chinese Journal of Mechanical Engineering, 2012, 25(6): 1287–1297
12. Leng Y.G., Wang T.Y., Guo Y., et al, "Engineering signal processing based on bistable stochastic resonance", Mechanical Systems and Signal Processing, 2007, 21:138–150
13. Tan J.Y., Chen X.F., Wang J.Y., et al, "Study of frequency-shifted and re-scaling stochastic resonance and its application to fault diagnosis", Mechanical Systems and Signal Processing, 2009, 23:811–822
14. Li Q., Wang T.Y., Leng Y.G., et al, "Engineering signal processing based on adaptive step-changed stochastic resonance", Mechanical Systems and Signal Processing, 2007, 21:2267–2279
15. Li B., Li J.M., He Z.J., "Fault feature enhancement of gearbox in combined machining center by using adaptive cascade stochastic resonance", Sci China Tech Sci, 2011, 54:3203–3210
16. Duan F. B., Chapeau-Blondeau F., Abbott D., "Stochastic resonance in a parallel array of nonlinear dynamical elements", Phys Lett A, 2008, 372:2159–2166
17. McDonnell M.D., Abbott D., Pearce C.E.M., "An analysis of noise enhanced information transmission in an array of comparators", Microelectron J, 2002, 33: 1079–1089
18. Stocks N.G., "Information transmission in parallel threshold arrays: suprathreshold stochastic resonance", Phys Rev E, 2001, 63:041114
19. Rousseau D., Chapeau-Blondeau F., "Suprathreshold stochastic resonance and signal-to-noise ratio improvement in arrays of comparators", Phys Lett A, 2004, 321:280–290
20. He Q.B., Wang J., "Effects of multiscale noise tuning on stochastic resonance for weak signal detection", Digit Signal Process, 2012, 22:614–621
21. He Q.B., Wang J., Liu Y.B., et al, "Multiscale noise tuning of stochastic resonance for enhanced fault diagnosis in rotating machines", Mechanical Systems and Signal Processing, 2012, 28:443–457
22. Zhang X.F., Hu N.Q., Hu L., et al, "Stochastic resonance in multi-scale bistable array", Phys Lett A, 2013, 377:981–984
23. Marek F., Emil S., "Stochastic resonance: a chaotic dynamics approach", Physical Review E, 1996, 54(2):1298–1304
24. Zhang X.F., Hu N.Q., Hu L., Zhe C., "Multi-scale bistable stochastic resonance array: A novel weak signal detection method and application in machine fault diagnosis", SCIENCE CHINA Technological Sciences, 2013, 56(9):2115–2123